STUDIES IN
LANGUAGE
COMPANION
SERIES 240

The Documentarist Turn

From observable linguistic behaviour to typological generalizations

Edited by
Sonja Riesberg
Uta Reinöhl
Birgit Hellwig

JOHN BENJAMINS PUBLISHING COMPANY

The Documentarist Turn

Studies in Language Companion Series (SLCS)
ISSN 0165-7763

This series has been established as a companion series to the periodical *Studies in Language*.

For an overview of all books published in this series, please see *benjamins.com/catalog/slcs*

General Editor

Elly van Gelderen
Arizona State University

Founding Editor

Werner Abraham
University of Vienna / University of Munich

Editorial Board

Alexandra D'Arcy
University of Victoria

Tine Breban
The University of Manchester

William A. Croft
University of New Mexico

Östen Dahl
University of Stockholm

Gerrit J. Dimmendaal
University of Cologne

Marianne Hundt
University of Zurich

Ekkehard König
Free University of Berlin

Christian Lehmann
University of Erfurt

Elisabeth Leiss
University of Munich

Marianne Mithun
University of California, Santa Barbara

Heiko Narrog
Tohuku University

Johanna L. Wood
University of Aarhus

Debra Ziegeler
University of Paris III

Volume 240

The Documentarist Turn
From observable linguistic behaviour to typological generalizations
Edited by Sonja Riesberg, Uta Reinöhl and Birgit Hellwig

The Documentarist Turn

From observable linguistic behaviour to typological generalizations

Edited by

Sonja Riesberg
LaCiTO – CNRS

Uta Reinöhl
University of Freiburg

Birgit Hellwig
University of Cologne

John Benjamins Publishing Company
Amsterdam / Philadelphia

TM The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/slcs.240

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2026008329 (PRINT) / 2026008330 (E-BOOK)

ISBN 978 90 272 3400 1 (HB)
ISBN 978 90 272 4388 1 (E-BOOK)

© 2026 – John Benjamins B.V.
This e-book is Open Access under a CC BY-NC-ND 4.0 license.
https://creativecommons.org/licenses/by-nc-nd/4.0

This license permits reuse, distribution and reproduction in any medium for non-commercial purposes, provided that the original author(s) and source are credited. Derivative works may not be distributed without prior permission.

This work may contain content reproduced under license from third parties. Permission to reproduce this third-party content must be obtained from these third parties directly.

Permission for any reuse beyond the scope of this license must be obtained from John Benjamins Publishing Company, rights@benjamins.nl

John Benjamins Publishing Company · https://benjamins.com

This one is for you, Nikolaus!

Table of contents

Preface xi

CHAPTER 1. Nikolaus Himmelmann and the documentarist turn 1
Birgit Hellwig, Uta Reinöhl & Sonja Riesberg

PART I. Research context

CHAPTER 2. The development of language documentation in Indonesia 28
Luh Anik Mayani & Katubi

CHAPTER 3. Collaborative research: Language documentation and the documentarist turn in Western New Guinea 45
Yusuf Sawaki, Hugo Warami, Apriani Arilaha, Jeanete Lekeneny & Boas Wabia

CHAPTER 4. Diversifying science: Insights of native- and non-native-speaker collaborations in linguistic descriptions 75
I Wayan Arka

CHAPTER 5. From the audible to the meaningful: The role of listener/speakers in making sense of 60 years of Dalabon recordings 107
Nicholas Evans & Manuel Pamkal

CHAPTER 6. Language archives in context 134
Felix Rau

CHAPTER 7. Science in public: The representation of *language documentation* in science communication 149
Tobias-Alexander Herrmann, Kai Kaspar & Stephan Packard

PART II. Documentary corpora

CHAPTER 8. Taking language use seriously: On the importance of multimodality and interaction in documentary practices 182
Mandana Seyfeddinipur, Pamela Perniss & Claudia Wegener

CHAPTER 9. The multimodal turn in corpus building: Lessons from language documentation and sign language corpora **210**
Anastasia Bauer, Sonja Gipper, Tobias-Alexander Herrmann, Jana Hosemann & Simona Sbranna

CHAPTER 10. Multimodal communication in historical perspective: Gestures and visual prominence in Pharaonic Egyptian monumental inscriptions **239**
Silvia Kutscher

CHAPTER 11. Reflections on literacy in language documentation **263**
Melanie Schippling

CHAPTER 12. More diachronic than you think: Historical depth within language documentation corpora and its potential to mitigate two major biases in linguistics **287**
Uta Reinöhl

CHAPTER 13. Songs in Eipo, Yale, Awiakay and Meakambut, four Papuan languages **314**
Volker Heeschen & Darja Hoenigman

CHAPTER 14. Transcription practices and (the documentation of) metalinguistic knowledge **343**
Katherine Walker, Birgit Hellwig & Dagmar Jung

PART III. Synchronic perspectives

Phonetics, phonology, and prosody

CHAPTER 15. Pauses, parts of speech, and word order: A comparative corpus study on 27 languages **376**
Frank Seifart

CHAPTER 16. Intonation contour length as a quantification of macro-rhythm **397**
Constantijn Kaland

CHAPTER 17. The phonetics and phonological behaviour of vowel length in Meto **429**
Kirsten Culhane & Owen Edwards

CHAPTER 18. Prosodic effects on glottal stop realization and voice quality variation in a corpus study of Tahitian **464**
Janet Fletcher & Adele Gregory

Morphosyntax

CHAPTER 19. High animacy does not favour zero over pronouns — especially for objects **488**
Geoffrey Haig & Stefan Schnell

CHAPTER 20. A global view on the *Ausrichtungsaffix* and the deep differences between nominalization-derived versus applicative-derived diathesis in Austronesian **518**
Daniel Kaufman

CHAPTER 21. What's a passive in Besemah? **572**
Bradley McDonnell

CHAPTER 22. The acquisition of symmetrical voice systems: Comparative notes on Tagalog, Indonesian, and Totoli **610**
Maria Bardají, Rowena Garcia, Christoph Bracks & Evan Kidd

Discourse structure

CHAPTER 23. On syntactic uniformity and the functional effects of verbal DPs **644**
Katharina Haude

CHAPTER 24. Information structure and the agent-first preference: A corpus-based case study of a free word order language **671**
Eva Schultze-Berndt

CHAPTER 25. Bridging demonstratives in recognitional use **704**
Petra B. Schumacher & Klaus von Heusinger

CHAPTER 26. Hierarchical discourse structure in spoken narratives **724**
Isabel Compes, Jakob Egetenmeyer & Katja Hannß

PART IV. Diachronic perspectives

CHAPTER 27. Language change in real time: Investigating conversational priming in repetitional responses **756**
Sonja Gipper, Eugen Hill & Pascal Coenen

CHAPTER 28. On similarity, mistake, and belief revision: The case of *yimarne(k)* in Kunbarlang **786**
Ivan Kapitonov & Caroline Gentens

CHAPTER 29. Grammaticisation of medial speech verbs: A case study from Yali (West-New Guinea) **815**
Katharina Gayler & Sonja Riesberg

CHAPTER 30. Of Malays, Moluccans and Papuans: The development and spread of Malay in Papua **844**
Antoinette Schapper

CHAPTER 31. On the origins of Javanese negators **866**
Alexander Adelaar

Epilogue

CHAPTER 32. Epilogue: *Episteme* and *techne* **900**
Stephen C. Levinson

Language index **903**

Subject index **907**

Preface

This volume celebrates one of the most creative, independent-minded and influential scholars in general linguistics, typology and the research of understudied and endangered languages today: Nikolaus P. Himmelmann. His contributions cover a wide range of topics and subfields, but if there is one thing that unites them, it is seeking an answer to the question of — in his own words — what it means to "take spoken language seriously". This question is at the heart of language documentation as he conceives it, and the increasing availability of well-curated, audiovisual documentary corpora for a constantly growing number of languages is now bringing us into the realm where we can start to address this question in earnest. This is the point of origin for this volume, as the contributors approach this question from various angles and explore different facets of how the documentarist turn impacts our way of thinking about language. In this way, Nikolaus himself has set the theme for this volume, which we dedicate to him.

We, the editors of this book, were socialized academically into the language documentation paradigm between the time when it was just emerging (BH starting her PhD in the year of Himmelmann 1998) and by the time it had become firmly established (SR and UR starting their PhDs around ten years into the programme). Nikolaus has been — and continues to be — deeply influential for our intellectual development and our ways of doing linguistics. More than that, he supported and guided us in navigating the academic minefield responsibly and ethically — sometimes in response to our requests, but more often than not by setting an example. It is his academic and personal integrity that has had a profound impact on us and shaped our own trajectories, and that we hope to emulate in our own encounters. No matter how seriously we take language in any of its modalities, we do not seem to find the right words to express the extent of our gratitude.

It seems we are not the only ones who were inspired by Nikolaus. Throughout the making of this volume, we have been moved by the continued enthusiasm and responsiveness of authors and reviewers alike.

Our sincere thanks go to all authors for creating this volume with us. Many thanks also for helping us with the internal reviewing process.

For their external reviewing, we are very grateful to Felix Ameka, Laura Arnold, Maria Averintseva-Klisch, Laura Becker, Axel Bohmann, Cat Butz, Alec Coupe, Sandra Debreslioska, Kristina Gallego, David Felipe Guerrero-Beltran, Charlotte Hemmings, Eva Huber, Hae-Song Jeon, Masha Khachaturyan, Paul Kroeger, Bruno

Olsson, Naomi Peck, Simon Rössig, Christoph Rühlemann, Fahime Same, Aung Si, Edgar Suter, Eva van Lier, Jozina Vander Klok, Jean-Christophe Verstraete, Yanti, and three reviewers who preferred to remain anonymous.

Our Cologne student assistants Beyhan Aliy, Anna Czech, Oriane Huet, Linus Kohler, Jacob Menschel, Lara Meuser, Alexander Poeplau, Ruben Rau, Katja Wiesner, and Mitzi Weiser have done a great job in helping us format and edit the chapters, including the tedious task of cross-checking the references. Many thanks to all of you!

We also thank Nick Evans for coining the term "documentarist turn", Mary Chambers and Katherine Walker for selective proofreading, and, once more, Aung Si for performing miracles without wanting to be named.

Finally, a special thanks to Elly van Gelderen, Ymke Verploegen and the team at the Studies in Language Companion Series for their exceptional support and their smooth handling of the entire process.

The publication of this volume as open access was supported by the Open Access Publication Fund of the University of Cologne.

Editors are listed in reverse alphabetical order, reflecting the length of our collaboration with Nikolaus.

Sonja Riesberg, Villejuif
Uta Reinöhl, Freiburg
Birgit Hellwig, Köln

CHAPTER 1

Nikolaus Himmelmann and the documentarist turn*

Birgit Hellwig[1], Uta Reinöhl[2] & Sonja Riesberg[1,3]
[1] University of Cologne | [2] University of Freiburg | [3] CNRS — LaCiTO

1. Introduction

The inception of language documentation can be traced back to the 1990s, when it began to emerge from earlier fieldwork-based descriptive and typological research traditions. Although clearly indebted to these traditions, language documentation did not merely continue them in a new guise; rather, it was a novel approach that brought data into sharp focus and advanced our theorizing about data. We now think of language documentation as a subdiscipline of linguistics, and indeed, the early years were marked by language documentation carving out its place within linguistics, testing its foundations empirically, and, in the process, fleshing out its theoretical and methodological agenda and developing more and more facets of its programme. However, its lasting and remarkable impact is not due to establishing yet another subdiscipline, but to addressing — from a documentary perspective — some of the central themes within general linguistics: the relevance of linguistic diversity for theory building at every level of language, communication, and cognition; an appraisal of the implications of different types of data and contexts of data production and reproduction; as well as the affordances of everyday communicative interaction, and the interplay between modality, level of spontaneity, and language structure. It is these fundamental issues that have come to the fore through the "documentarist turn"; issues that are relevant not only to language documentation, but also to the field of linguistics as a whole and to the humanities and cognitive sciences at large. Today, language documentation is an integral part of the academic landscape and it is difficult to imagine contemporary linguistics without it.

A lot has been achieved since the 1990s, when the magnitude of language endangerment was registered, the plea for a "responsible linguistics" (Hale et al.

* With the help of Isabel Compes, Carmen Dawuda, Dagmar Jung, Felix Rau, Fritz Serzisko, Vera Szöllösi-Brenig, Katherine Walker, Helga Weyerts-Schweda & Lena Wolberg.

https://doi.org/10.1075/slcs.240.01hel
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

1992: 1, Editor's note) articulated, and the theoretical and methodological foundations laid down in Himmelmann's (1998) seminal paper *Documentary and descriptive linguistics*. The contributors to this volume map out the impact of the documentarist turn, and demonstrate how it shapes our way of thinking about language and doing linguistics. In this introduction, we take a step back and look at salient developments in Europe and, in particular, in Germany that underpin these advances. Language documentation was not invented by a single person, but has evolved over time through the interaction of many scholars and community members, all contributing their unique perspectives against a backdrop of various funding, university, and archiving infrastructures. Language documentation is very fortunate to be sustained by a large and active documentarist community. Our intention is *not* to write an encompassing history of language documentation, but to approach it through the lens of one of its central players: Nikolaus Himmelmann, to whom we dedicate this volume. Section 2 sets the scene with a brief introduction to language documentation, Section 3 focusses on Himmelmann's contributions and the Cologne academic environment, Section 4 addresses his role in building structures and infrastructures, and Section 5 gives an outline of this book.

2. Setting the scene: Language documentation

In January 2019, UNESCO proclaimed the period from 2022 to 2032 to be the *International Decade of Indigenous Languages*, with the goal "to draw attention to the critical loss of indigenous languages".[1] To many linguists today, language endangerment and the accelerating speed of language death that we have been witnessing since the 20th century is a topic of major concern. But for a long time there was only little awareness that the world is losing large parts of its linguistic diversity and heritage. A major trigger for change was a presentation by Johannes Bechert at the *14th International Congress of Linguists* in East Berlin 1987 (published as Bechert 1990). In his talk, Bechert, citing Mühlhäusler (1985), compared language loss — and its consequences — to the extinction of plant and animal species, and pointed to the fact that, unlike scientists in the field of biology, linguists have been surprisingly idle in addressing the problem. The reaction to this wake-up call was immediate (see Himmelmann 2008a): the general assembly of the *Comité International Permanent des Linguistes* passed a resolution to recognize endangered linguistic diversity as its central concern, and the topic spread rapidly as a matter of urgent and active discussion in the worldwide linguistic

1. A/RES/74/135 General Assembly.

community. The most pressing question was how to conceptualize and structure a good record of a (disappearing) language such that it is maximally useful for multiple purposes — from linguistic and other scientific research, to fulfilling the goals of language communities and the broader public — both now and in the future.

The answer to this question grew out of the long-established traditions and practices of descriptive linguistics and fieldwork-based research. Field recordings have been made and stored since the late 19th century, some of which are nowadays accessible in digital archives and databases. *The Archives for Traditional Music*, for example, hosts recordings by Franz Boas and Edward Sapir, and audio recordings from the early 20th century can be found across many contemporary archives, alongside other legacy materials. Particularly in the US and in Australia, writing a descriptive grammar of an un(der)described language as a PhD thesis has long been not only accepted but, in certain places, even encouraged; the "Boasian trilogy" (Evans & Dench 2006: 10) to which a grammar belongs also includes a lexicon and a text collection, providing a data basis for, and to some extent, facilitating the replication of descriptive statements. It has obviously also been the case that, long before the rise of documentary linguistics, field linguists have valued community involvement and collaborative research with community members, and have reflected on ethical questions (e.g. Samarin 1967; Bouquiaux & Thomas 1976). Likewise, reflections on data collection methods and the naturalness of different data types were nothing new. Boas (1917: 1) notes that "an unnatural simplicity of syntax prevails in most of the dictated texts", and it was said of Bloomfield that he "would never elicit paradigms, for fear he would induce his informant to say something under the artificial conditions of talking with an outsider that he would never have said in his own everyday surroundings" (Hockett 1968: 89–90).

But while language documentation as the response to the loss of linguistic diversity *builds on* existing traditions, its rise represented a sea change, as it allowed us to consider all of the above-mentioned aspects together. Himmelmann's (1998) ideas were ground-breaking, not because he advocated for an intensification of fieldwork or for making field recordings accessible in a more systematic and long-lasting way, but because he pled for a strict conceptual separation of documentary and descriptive activities, demanding that "the collection and presentation of primary data [should] receive the theoretical and practical attention they deserve" (Himmelmann 1998: 164). In other words, he requested that we "take spoken language seriously". With respect to grammar writing, Evans and Dench later called this the "naturalistic turn" (2006: 25): with language documentation becoming common practice, grammatical descriptions were no longer based only on elicitation and a (frequently small) corpus of narrative texts, but on more, and more ecologically valid, data types. We call it the "documentarist turn" to highlight the fact

that documentary approaches now resonate across the entire spectrum of linguistic practices, and — above all — to capture what we see as the central achievement of language documentation: to have changed linguistics as a field, from advancing the practice of grammar writing, to actively shaping the linguistic research landscape as a whole.

3. The Cologne academic biotope

The early wake-up calls sounded by Bechert (1990) and Hale et al. (1992) did not go unheeded. Linguistic departments and community organizations around the world set out to develop academic and social responses to the newly recognized threat to linguistic diversity and the increasing loss of everyday linguistic practices and experiences. From the very beginning, the University of Cologne established itself as one of the main centres for these activities. It is within this context, with its many compelling personalities and inspiring encounters, that we must view Nikolaus's work and his influential ideas that would go on to shape language documentation for years to come.

Nikolaus's early career began, in fact, not in Cologne, but at the Ludwig-Maximilians University of Munich (LMU Munich) and at the University of California in Los Angeles (UCLA). His MA thesis on *Linking im Tagalog* [Linking in Tagalog] (1983) laid the foundations for a rich, ongoing research programme on Austronesian languages. And his PhD thesis on *Morphosyntax und Morphologie: Die Ausrichtungsaffixe im Tagalog* [Morphosyntax and morphology: The orientation affixes in Tagalog] (1987) sparked a lifelong interest in the symmetrical voice systems of Philippine-type languages (e.g. Himmelmann 2002a, 2002b, 2008b; Himmelmann & Riesberg 2013; Riesberg et al. 2022; Bonmann et al. 2024). His interest in Austronesian languages also shines through — at least as strongly — in his documentary activities, which include documentation projects on several Austronesian languages of Indonesia and East Timor (especially Waima'a, Totoli, and Wooi), and later on Papuan languages of the region (especially Iha).

His close and long-standing relationship with the Cologne linguistics department began shortly after he completed his MA studies. He joined the department in 1984 as a researcher in Hansjakob Seiler's UNITYP project,[2] one of the earliest fully fledged theoretical paradigms combining typological observations with func-

2. UNITYP: *Sprachliche Universalienforschung und Typologie unter besonderer Berücksichtigung funktionaler Aspekte* [Research into linguistic universals and typology with special consideration of functional aspects], see UNITYP (https://ifl.phil-fak.uni-koeln.de/prof-em-seiler/persoenliche-seiten/unityp).

tionalist explanations and a focus on un(der)studied languages. The department was rapidly becoming a hub for fieldwork-based research, including Hansjakob Seiler's fieldwork on Cahuilla, Fritz Serzisko's research on Ik, and, later, Hans-Jürgen Sasse's work on Cushitic, Semitic, Iroquoian, and Balkan languages. It is in this environment that Nikolaus's first publication was conceived (Himmelmann 1986), and these traditions of thought have remained deeply influential on his work and on subsequent generations of linguists at Cologne.

After his early stint as a researcher in UNITYP, Nikolaus, having received his PhD in 1986 from LMU Munich, returned to work at the Cologne department from 1989–1996 as an assistant professor, a phase in his career which culminated in 1997 in his "Habilitationsschrift" [postdoctoral thesis], *Deiktikon, Artikel, Nominalphrase: Zur Emergenz syntaktischer Struktur* [Deictic, article, noun phrase: On the emergence of syntactic structure]. His close collaboration with Hans-Jürgen Sasse was probably one of the most stimulating of his early career. Like Hans-Jürgen, Nikolaus would develop a deep commitment to fieldwork and a profound concern in the face of accelerating language death (Sasse 1990). The next stage of his career took him to the Australian National University (1996–1998), where he spent two happy and inspiring years. He briefly returned to Cologne as a Heisenberg scholar (1998–1999), followed by his first professorship in General Linguistics at the University of Bochum (1999–2007). He left Bochum for the University of Münster (2007–2010), before returning to Cologne for good in 2010. The fact that, after his return from Australia, his primary residence was always in Cologne seems in retrospect something of a give-away.

Throughout the 1980s and 1990s, Nikolaus formed long-lasting intellectual bonds. In Munich, he attended classes by Hans-Jürgen Sasse and Ulrike Mosel; in Los Angeles, he studied with Sandra Thompson and Paul Schachter; in Canberra, he was looked after by Andy and Medina Pawley; and in Cologne, besides his interactions with Hansjakob Seiler and Hans-Jürgen Sasse, it was especially the many — or, in Nikolaus's words, the one never-ending[3] — conversation(s) with Fritz Serzisko, with his interest and expertise in spoken language and discourse structure (e.g. Serzisko 1992), that left a lasting impression. Conversations with Thomas Müller-Bardey shaped Nikolaus's ideas about grammaticalization, and he shared an interest in Tagalog — but also in Kölsch (the Cologne dialect) — with Werner Drossard. Students from the time recall a "breath of fresh air" and remember vividly their coursework based on "real" fieldwork data as well as Nikolaus's "critical, but always helpful" feedback. This was also a time of legendary

3. According to institutional memory, he is said to have stated: "mit Fritz habe ich angefangen zu reden und nicht wieder aufgehört [I started talking with Fritz and we have never stopped since]".

parties that remain firmly anchored in the department's memory and (sadly, non-archived) photographic record.

This environment proved profoundly influential, and provided perhaps the ideal conditions for Nikolaus's conception of language documentation to grow and flourish. Here, he continued and refined his research programme in the affordances of speech-act and discourse contexts and their impact on diachronic and synchronic linguistic structure — a foundation that underlies modern language documentation and connects it to his diverse academic interests.

As a grammaticalization researcher, Nikolaus continued Joan Bybee's work to highlight the constructional context within which, and with which, a gram(maticalizing morpheme) develops. His postdoctoral thesis — *Deiktikon, Artikel, Nominalphrase* (1997) — remains a touchstone in the literature on morphosyntactic change. The work takes a holistic approach to grammaticalization, elucidating ways in which phrase structure emerges alongside the grammaticalization of its central element, the gram. It is in this work, too, that Nikolaus's persistent interest in deixis feeds into his research into emerging grammatical structure. In *Deiktikon*, in his influential paper *Gram, construction, and class formation* (2005), and elsewhere, Nikolaus's work has become particularly well known for his clarification of host class expansion as both a key by-product of grammaticalizational changes and a central force driving word class formation and change. *Lexicalization and grammaticization: Opposite or orthogonal?* (2004) remains a central paper delineating and disentangling the relationship between grammaticalization — or grammaticization as he prefers to call it, for brevity, style, and the affiliation with Bybee's work — and other types of language change, a line of work continued in *Renewal: A figure of speech or a process sui generis?* (Reinöhl & Himmelmann 2017).

His work on grammaticalization in combination with his intense engagement with spoken data have led Nikolaus to develop a long-standing interest in prosody and the prosody-syntax interface. He has surveyed the prosodic systems of Austronesian languages (e.g. Himmelmann & Kaufmann 2020; Riesberg et al. 2020), produced important contributions on the typology of stress (Himmelmann 2022b), and highlighted the role of prosodic chunking in creating the types of linguistic units in which grammaticalization may arise (Himmelmann 2022a).

As a typologist, Nikolaus combines his dedication to fieldwork and spoken language in use with a keen interest in mapping the variation space of human language(s). In addition to the phenomena already mentioned, including voice systems and stress, Nikolaus has provided, among other things, typologies of demonstratives (Himmelmann 1996), secondary predication (Schultze-Berndt & Himmelmann 2004; Himmelmann & Schultze-Berndt 2005), and word-order regularities (Riesberg et al. 2019). Over the past decade, he has collaborated with Cologne-based colleagues, including Beatrice Primus, Klaus von Heusinger, and

others, to develop a theoretical architecture of prominence phenomena (Himmelmann & Primus 2015). Throughout his career, Nikolaus has not only made significant contributions to typologizing realms of linguistic structure, but has also provided the field with meta reflections on ways of typologizing (Himmelmann 2000) and the standing challenge of language comparison (Himmelmann 2022c).

The above — incomplete — list testifies to Nikolaus's wide-ranging intellectual curiosity. Yet, his varied academic interests are united by a common thread: what it means — for all facets of language — to "take spoken language seriously". This is the theme that has guided Nikolaus's career, and it is reflected in his conceptualization of language documentation as a research paradigm capable of making lasting contributions to linguistic research and language communities; contributions that transcend individual research concerns and questions. Building on the foundations in Himmelmann (1998, *Documentary and descriptive linguistics*) and Himmelmann (2006a, *Language documentation: What is it and what is it good for?*), Nikolaus made important contributions to theorizing about data types (e.g. Himmelmann 2012); understanding practices, styles, and factors influencing transcription (e.g. Jung & Himmelmann 2011); and productively adapting concepts and practices from Conversation Analysis-style approaches to the specific purposes of language documentation with a special focus on prosody, supported by a Research Professorship for Prosodic Documentation (Volkswagen Foundation, 2004–2005) (e.g. Himmelmann 2006b, 2006c; Himmelmann & Ladd 2008). He has continued to shape the discipline and to reflect on its role in, and in interaction with, typology and cross-linguistic theorizing (e.g. Himmelmann 2012, 2022c). Throughout, his conception of language documentation has been embedded within a broader societal, academic, and infrastructural context (see Section 4).

Nikolaus's many achievements have been formally recognized in the field, from the *Best Paper in Language* 2014 award, to his election as an Honorary Member of the Linguistic Society of America (2018), to the special publication *Reflections on language documentation 20 years after Himmelmann 1998* (McDonnell et al. 2018), to national honours (e.g. his election as member of the review board for linguistics at the German Research Foundation (DFG); see Section 4.1) and local awards (e.g. *Leo-Spitzer Preis* 2017 of the University of Cologne). The DFG has honoured Nikolaus's lifetime achievement with a *Reinhart Koselleck-Projekt* — a project that is awarded to "outstanding researchers with a proven scientific track record to pursue exceptionally innovative, higher-risk projects."[4] He is one of only three linguists to have received this award. The title of this project is programmatic and highly fitting: *Gesprochene Sprache ernst nehmen* [Taking spoken language

4. DFG, German Research Foundation — Reinhart Koselleck Projects.

seriously].[5] As will become clear in the remaining sections of this introduction and in the chapters to follow, we have adopted it as the central theme running through this book.

4. Building structures

Language documentation developed — and continues to develop — not in a vacuum, but in the context of academic and societal preoccupations, affordances, and structures. Science foundations have played a crucial role in this respect, incentivizing documentation efforts and putting language documentation on the agenda (Section 4.1). This goes hand in hand with an institutionalization of language documentation concepts and practices within university departments, teaching curricula, and doctoral programmes (Section 4.2), as well as with efforts to diversify academia (Section 4.3). From the beginning, issues of data curation and availability have been a central concern, allowing language documentation to become a driving force for creating a sustainable research data and archiving infrastructure in the humanities (Section 4.4). Finally, documentary research reaches well beyond academia and is set within a broader societal discourse on the relevance and promotion of linguistic diversity (Section 4.5).

These structure-building aspects of language documentation do not always receive the recognition they deserve, yet it is precisely these aspects that put language documentation on the agenda and ensure its long-term sustainability. For the ongoing success of language documentation, it is fortunate that enough people had the strategic foresight to recognize the importance of these structures early on and have worked ever since on creating, maintaining, and improving them. In this section, we trace developments in relation to Nikolaus's activities: in his vision, structure-building is not relegated to the sidelines, but forms a central part of language documentation. Consequently, many of his activities have impacted language documentation more generally and are relevant beyond Germany. The topics in 4.3, 4.4, and 4.5 are taken up again in Part I of this volume (see Section 5), and we therefore provide only brief sketches in this introduction; topics in 4.1 and 4.2 are presented in more detail.

5. DFG — GEPRIS — Gesprochene Sprache ernst nehmen (https://gepris.dfg.de/gepris/projekt/523828585?context=projekt&task=showDetail&id=523828585&).

4.1 Putting language documentation on the agenda: The role of science foundations

The rise of language documentation is inextricably linked to the development of accompanying structures in research funding. In the late 1990s, non-profit initiatives started to offer small-scale project funding (see Section 4.5), and research funding quickly gained momentum thanks to two dedicated large-scale funding schemes: *Dokumentation bedrohter Sprachen* [Documentation of Endangered Languages] (DobeS, 2000–2013, Volkswagen Foundation) and the *Endangered Languages Documentation Programme* (ELDP, since 2002, Arcadia Fund). Subsequently, national science foundations also came on board, and today language documentation has become firmly established in the funding landscape, as funding can be applied for under the usual programmes of national and international science foundations, e.g. of the Deutsche Forschungsgemeinschaft [German Research Foundation] (DFG).

In their ability to promote research activities, science foundations played a major role in creating an academic community of documentarists, and have continued to support them in advancing and renewing the methodological and theoretical framework of language documentation. Nikolaus was exceptionally active in efforts to get documentation on the agenda of funding bodies: he played a pivotal role at DobeS, and was a long-standing elected member in the DFG review board for linguistics. In the following paragraphs, we outline developments through the lens of DobeS (as the first large-scale initiative) and the DFG (as an example of developments at a national level).

The DobeS programme was probably *the* initiative that set a world-wide process in motion and successfully anchored language documentation within the funding landscape. DobeS was established in June 1999, but the idea dates back to 1993 when the Volkswagen Foundation funded a summer school at the University of Cologne (see Section 4.5). The Foundation seized upon two transformations that were underway in the mid-1990s/2000s: one was the digital revolution; the other was the political recognition of the world's intangible cultural heritage by UNESCO, which led to the drafting of the *UNESCO Universal Declaration on Cultural Diversity* as a framework in 2001 and the *Convention for the Safeguarding of the Intangible Cultural Heritage* in 2003. Against this background, the aims of DobeS were to (1) promote the documentation of endangered languages, (2) further develop documentary linguistics on this basis, and (3) create a digital multimedia archive together with corresponding software tools.

DobeS started with a one-year pilot phase (eight exploratory documentation projects, one multimedia database project), continued with a main phase (numerous three-year projects), and ended with a final phase. In the final phase, an addi-

tional focus was set: exploring issues in the re-usability of archival documentary records, and thereby testing whether the aim of language documentation to create a "lasting, multipurpose record of a language" (Himmelmann 2006a: 1) has been fulfilled. Over a period of 14 years, DobeS was able to fund 103 projects — on around 100 languages — with a total of around € 28 million, and the DobeS archive became one of the first genuinely digital archives in the world. In June 2013, the initiative ended with an international conference entitled "Language Documentation: Past — Present — Future" at Herrenhausen Palace in Hanover, which was attended by 183 participants from 28 countries.

DobeS not only offered funding, but — perhaps more importantly — provided a framework that enabled people to interact and advance language documentation conceptually and methodologically. An important milestone in this respect was a 2003 summer school at the University of Frankfurt, which resulted in the first handbook of documentary linguistics, *Essentials of Language Documentation* (Gippert, Himmelmann & Mosel 2006). Not surprisingly, Nikolaus contributed the fundamental conceptual first chapter *Language documentation: What is it and what is it good for?* The Foundation provided a publication grant so that the book would be affordable to students, early career researchers, and researchers from the global south.

The 2013 closing conference was co-organized by Nikolaus, together with Nick Evans (Australian National University), Ulrike Mosel (University of Kiel), and Mandana Seyfeddinipur (ELDP). With DobeS coming to an end, one point of discussion was how to secure future funding. Under Nikolaus's leadership, a memorandum was drafted, and the Volkswagen Foundation forwarded it to all central policy-makers within the German funding system — the DFG, the Federal Ministry of Education and Research, the Academies of Science and Humanities, among others — urging them to fund language documentation activities as part of their regular funding programmes.

Vera Szöllösi-Brenig, the programme officer at the Volkswagen Foundation responsible for DobeS, describes her experience as follows:

> I feel very fortunate to have been responsible for DobeS from the drafting of the very first leaflet in 1999 to the final conference in 2013. I was equally fortunate to have met and worked with such interesting people as Nikolaus: people who not only conduct research at their desks and in libraries, but whose workplace is "the world". As for Nikolaus, I got to know him as an excellent linguist with high levels of expertise and life skills. And not to forget: a committed and critical but constructive person I could always rely on. As Nikolaus wrote in the book *14 Jahre DobeS* [14 Years of DobeS] about the many individuals involved in DobeS: "In the end, it is the people who make a programme a success."

Through its strategic allocation of funds, DobeS helped establish a documentarist community, advance the field conceptually and methodologically, and prove the scientific and societal value of language documentation. It successfully put language documentation on the agenda — to the extent that the end of DobeS did not mean the end of funding.

In Germany, the DFG — as the largest German research funding organization — had already started to take note of the emerging language documentation paradigm when, in October 2007, the US National Science Foundation organized an international workshop to discuss language documentation strategies and facilitate networking. Susanne Anschütz, the programme officer for linguistics at the DFG, participated in that meeting and enthusiastically reported back to her colleagues. Only a month later, the new programme officer, Helga Weyerts-Schweda, would go on her very first business trip to the European Science Foundation in Brussels, where a new funding programme — *Better Analyses Based on Endangered Languages* (EuroBABEL) — was being put together. Again, Nikolaus assumed a leading role as co-initiator, together with colleagues from across Europe. The goal was to take language documentation one step further, and to bring "the newly gathered data to bear on the development of linguistic theory and all areas concerned with the study of language."[6]

From the perspective of the DFG, this proposal contained some highly attractive elements: the important topic of linguistic diversity, the bridge between language documentation and linguistic theory, the opportunity to cooperate with European partners, and — thanks to the success of DobeS — the strong anchoring of documentary linguistics in the German academic landscape with its promise of excellent project applications. The DFG inaugurated the programme in 2008, and 20 national funding bodies teamed up to fund 11 large-scale cross-national EuroBABEL projects over a period of three years (2009–2011).

Unlike the Volkswagen Foundation, the DFG does not (as a rule) set thematic priorities, and funding within its standard programmes is not orientated towards documentary activities in their own right. Emphasis is instead placed on the uses of data — documentary or otherwise — for linguistic analyses and research questions. EuroBABEL was an important milestone in this respect, as it directly targeted the relevance of language documentation for theory building. Though the DFG had already funded fieldwork-based projects that met these standards before EuroBABEL, in subsequent years the number of such projects grew: the documentarist turn led to an increase in high-quality proposals that brought together language documentation and linguistic theory. Furthermore, the increas-

6. EuroBABEL: European Science Foundation (http://archives.esf.org/coordinating-research/eurocores/programmes/eurobabel.html).

ing importance of language documentation came to be reflected in the structure of the review board. In the DFG system, review boards are elected every four years by the entire academic community — its members are selected in a competitive process, gaining the trust of their fellow researchers. Twice, Nikolaus was elected to the board for linguistics, serving the maximum of two periods (2016–2019, 2020–2023); in the second period, the board elected him as their spokesperson. Anecdotally, some of his contributions seem to have outlived his term in office: his unofficial rule that "projects that are twice as expensive as normal must also be twice as good" is apparently still being passed on to new members of the board.

4.2 Growing up: Language documentation and the university system

An unmistakable sign of a successful research paradigm is when it becomes institutionalized within the university system. This is what we observe in the development of language documentation: departments started to engage with the emerging programme, documentarists were hired and tenured, and documentary concepts and methods found their way into teaching curricula. Not surprisingly, Nikolaus actively encouraged these developments at his academic institutions. On a basic level, this entails the obvious: bringing together a critical mass of researchers in one place, as well as training and supporting — in Nikolaus's words — *diese jungen Leute* [these young people]. One way to do this is to take advantage of the funding and employment opportunities offered by science foundations. Nikolaus has made the most of these opportunities, designing and consistently attracting funding for large-scale projects, as well as encouraging and supporting others throughout the application process. But third-party funding is only one — albeit important — step in a long process. Its temporary nature means that universities eventually have to buy in and make at least some arrangements permanent. This, in turn, can only be achieved through administrative and strategic meetings and discussions at all levels of the university hierarchy — a thankless task, often frustrating, and usually invisible. Language documentation can count itself lucky that a good number of documentarists did not, and do not, shy away from this. In Cologne, Nikolaus endlessly negotiated with a view to improving research conditions, establishing a sustainable research infrastructure, and working towards turning externally funded, temporary positions into tenure-track positions.

The most visible result of these efforts is a large number of researchers affiliated with the Cologne department at any given time — BA/MA research assistants gaining their first experience as members of project teams, PhDs and junior postdocs being guided in developing their own profiles, senior postdocs taking their first steps in coordinating large projects, established researchers benefiting from discussions and strategic advice, and, in the midst of all this, visiting researchers

and collaborators from language communities from all over the world contributing their unique perspectives and experiences. Remarkably — because it is not a given — this also applies to the administrative perspective. Nikolaus put "research management" on the agenda long before it became a hot topic: administrative processes were established and standardized; administrative work became more demanding but also more exciting and varied; and managing the numerous activities, projects and people — none of which are alike — offered considerable scope for creativity within routine tasks.

Building such a community is no mean feat, but maintaining it is even more of a challenge. On a personal level, it requires taking individual needs and aspirations seriously and advising students, researchers, and administrators with their best possible interests in mind — even in cases where these might conflict with one's own. On an intellectual level, it requires curiosity and openness to new ideas and people in order to foster and maintain the vibrancy of the community. Nikolaus can take full credit for having achieved all this. He seeks out people at all career levels whose areas of expertise and perspectives overlap with, or differ from, his own, and who can build bridges between language documentation and any other field. Throughout his career, he has supervised, mentored, and advised people with very different approaches: diachronic or synchronic specializations, corpus or experimentally orientated methods, descriptive or formalist approaches to language, work on un(der)studied languages or on major European languages, and much more. Conversely, he encourages people — from students through to experienced researchers — to leave Cologne, to broaden their horizons and gain experience, maintaining contact all the while. The impact of these practices can be seen not least in this volume, as editors and contributors alike are all connected in one way or another to Cologne — some having started their careers here, others arriving at later stages; some leaving, others returning; some passing through, and others staying.

There is a broader vision behind these comings and goings, which brings us back to the institutionalization of language documentation within the university landscape: to place it at the very heart of doing linguistics and thinking about linguistics. This perspective on language is reflected in Nikolaus's continuous efforts to build structures and bridge (sub-)disciplines, as he proactively seeks out colleagues from across the university with the goal of bringing together different academic perspectives on fundamental issues.

He thus became one of the driving forces behind the never-ending process of institutionalizing and constantly renewing the language sciences at the University of Cologne. This includes especially the Cologne Center of Language Sciences (CCLS; founded in 2002), which brings together language-related fields from across the university and supports the development of collaborative initiatives,

including the large-scale Collaborative Research Center (CRC 1252) *Prominence in Language* (2017–2028). In 2019, the University of Cologne recognized these efforts and established *Skills and Structures in Language and Cognition* (SSLAC) as a university-wide focal area. Nikolaus was a key figure in all these endeavours, often assuming leadership roles: director of CCLS (2010–2012), central member of the core group that conceptualized and prepared the CRC 1252 (2014–2017) and its first co-spokesperson, and spokesperson of SSLAC (2019–2025).

Today, these institutions have become focal points at the University of Cologne and place language documentation within a whole collaborative network of specializations: historical-comparative linguistics, phonetics, discourse studies, a range of language and literature studies as well as German Sign Language, different fields within the cognitive sciences, computational and machine-learning approaches, multilingual language learning and teaching, media studies, and philosophy. As a result, language documentation transcends the narrow confines of a subfield and emerges as an integral part of linguistics in general. Several of the contributions to this volume originated in this context and bring together authors with very different perspectives and areas of expertise.

The same vision has found its way into Cologne's teaching curricula. These, of course, include specialized courses on language documentation theory and methods, but — possibly more importantly — general linguistics is taught through the lens of language documentation, as its central tenets permeate the entire curriculum. That is, students are exposed early on to the idea that language documentation is not a narrowly defined subfield of linguistics, but has something to offer to linguistics as a whole, and students learn to integrate language documentation perspectives with whichever other specializations they pursue. Again, Nikolaus provided the impetus to redesign the study programme. Here, he was able to build on his experiences at the University of Münster, where he was director of the MA programme and founding director of the *Graduate School Empirical and Applied Linguistics* (2008–2010). In Cologne, he initiated and guided discussions at departmental and cross-departmental levels, which eventually led to the implementation of new curricula and the formation of a cross-departmental MA programme as well as joint PhD classes (2012–2014). Throughout all this, he took an active part in teaching advanced BA/MA students and PhDs, including fulfilling the role of mentor at the *a.r.t.e.s. Graduate School for the Humanities* (2010–2014). But his special interest was in the introductory courses, which he spent many years teaching, reorganizing, and improving — and he is still entertaining the hope that, one day, he will find the time to turn the resulting contents into an introductory textbook for linguistics.

4.3 Diversifying the field

Language documentation originated as a scientific endeavour in the context of western academia, where there is a strong power imbalance: the global north dominates and sets the research agenda, while the global south is often relegated to the role of providing assistance and producing data. From its inception, language documentation thus had to seriously engage with sometimes starkly asymmetric relationships across the north-south and academia-community spectrum. The field is making considerable progress towards fostering more participatory models, institutionalizing collaborative structures, and furthering the visibility of previously more marginalized research — with the result that we are seeing progress in incorporating diverse perspectives and academic traditions into the formulation of research questions, the design of projects, and in authorship attribution (e.g. addressed in recent special publications of *Language Documentation and Conservation*: Cruz Cruz 2020; de Lima & Riestenberg 2020; Sandman et al. 2022; Stirtz et al. 2024).

These developments were foreshadowed by language documentation being, by its nature and ideals, an inclusive undertaking. With its overarching goal to create a record of a language that is lasting and multipurpose, it invites interdisciplinary collaborations as well as collaboration beyond academia. Linguists have teamed up with anthropologists, musicologists, biologists, (ethno-)botanists, and others, in order to be better able to capture the linguistic, cultural, and biological diversity they encounter. More importantly, the involvement and active participation of the language community in the documentary process is key to ethical and successful language documentation, and allows language experts and documentary linguists to learn from each other. This can also provide the initial impetus for individuals to embark on their own academic trajectory, thus diversifying the field and science in general. Offering training in basic techniques and practices of language documentation is often one of the first things a researcher does when they start a new project. A language user who starts out as a local collaborator might thereby become an independent language activist (see Arka, this volume), or enrol in a linguistics programme and pursue a career in linguistics. With the institutionalization of language documentation in university programmes, students in many parts of the world are encouraged to work on the documentation of their and their country's languages, many of them endangered and un(der)described. At the same time, collaborations between local and (foreign) national institutions and universities have been set up, fostering joint documentation projects and encouraging the co-supervision of students by linguists from diverse academic traditions.

One of Nikolaus's central concerns in this respect is to create opportunities for language community members and the students and scholars local to them. The "Bali workshops" he organized in 2006 and 2007 heralded the start of modern documentary linguistics in Indonesia (Riesberg, to appear; Mayani & Katubi, this volume) and have had a lasting impact: many of its participants have obtained PhDs in (documentary) linguistics — either in Indonesia or abroad — and are now actively shaping the Indonesian (linguistic) research landscape.

But probably the largest share of Nikolaus's fieldwork time, energy, and passion was dedicated to CELD — the Center for Endangered Languages Documentation at the University of Papua in Manokwari, West Papua (see Sawaki et al., this volume). CELD was established in 2009 with initial funding from one of Nikolaus's DobeS projects; its director, Yusuf Sawaki, being one of the Bali-workshop alumni. A substantial portion of Nikolaus's third-party funding and prize money from awards found its way to Manokwari. But more remarkable than his financial support for CELD is the time and energy Nikolaus invested. The hundreds of hours spent sitting with CELD staff discussing finances; the endless checking of ELAN and dictionary files together with native-speaker staff; the constant insistence that metadata, data management, and data backups are imperative; the many workshops, teach-ins, and reading circles for local students — none of this shows up in CVs or publications. Much of this work he could have easily delegated to his staff from Cologne. But sharing knowledge, taking the time to explain, and giving people the opportunity to learn constitutes the core of a true collaboration between equals. These activities that continued over many years have without doubt contributed substantially to the success and today's independence of CELD.

4.4 Archiving infrastructures

Within language documentation, the attention to primary data goes hand in hand with equally careful attention to the concepts and practicalities of research data management. Language archives have evolved to honour the foundational principle of language documentation: to create documentary records that are "lasting" and "multipurpose". To do so, they design and implement strategies for the long-term preservation of documentary data, as well as paying heed to issues of accessibility and re-usability for diverse user groups (see Evans & Pamkal, this volume, for an example of working with archival and legacy data). The relationship between documentation and archiving is reciprocal, such that developments on either side influence and enrich the other (see Rau, this volume). Following their institutionalization, language archives continued to develop in dialogue with language documentation research practices, and are nowadays increasingly tasked with facilitating participatory models of research, supporting the co-creation of

knowledge, and removing social, geographical, and technical access barriers — i.e. contributing to diversifying academia and engaging the public. Likewise, developments in language archiving cannot be viewed in isolation from developments in academia more generally, where issues of data accountability, replicability, and re-usability are rapidly gaining importance. Language archives are thus anchored within language documentation, but at the same time constitute bridges to larger national and international research data initiatives. Given their early orientation towards research data, language archives were able to anticipate and influence many future developments, and they continue to play active roles in creating sustainable research data infrastructures across the humanities.

Throughout his career, Nikolaus made major contributions to the development of a research infrastructure for language data. This goes back to the initial phase of DobeS, when he identified one of the central concerns: the indispensable — yet not always smooth — interaction between linguistics and computer science for the purpose of developing tools for analysis and archiving that would meet documentarists' needs. In Peter Wittenburg from the Max Planck Institute for Psycholinguistics in Nijmegen he found a like-minded partner from the computational and engineering side. As a result, DobeS workshops and training courses were held every year in Nijmegen from 2002 onwards, where linguists and computer scientists were brought together to negotiate this prickly relationship in detail: What tools were necessary? Were they feasible? Later, as a member of the DFG's review board for linguistics, Nikolaus took on the responsibility for research data management and expanded the preliminary guidelines developed by the previous board. Today, research data statements are an integral part of all applications, and the review boards pay particular attention to ensuring that proposals address this issue.

For many years, Nikolaus directed considerable effort towards institutionalizing archiving infrastructures and thus ensuring their continued existence. He was not only instrumental in securing substantial funding for infrastructure, software development, and networking projects, but he also consistently embedded them within a larger data management context, thereby establishing language documentation and archiving as a central player in faculty- and university-wide strategic developments. In this way, he successfully set up the Language Archive Cologne (LAC; since 2012) and established close communication with university bodies (especially the Data Center for the Humanities, DCH). He pushed for the DCH to be institutionalized as a permanent facility at the Faculty of Arts and Humanities, and for it to function as an interface between, on the one hand, the *National Research Data Infrastructure Germany* initiative, and, on the other hand, international initiatives across the language sciences, especially the CLARIN Knowledge Center for Linguistic Diversity and Language Documenta-

tion (CKLD) and the Digital Endangered Languages and Musics Archives Network (DELAMAN).

Here, too, Nikolaus's foresight is evident: language documentation is not isolated, but is instead firmly anchored within overarching developments in the academic landscape. He acted on this vision early on and continues to advocate integrating language archives into (digital) library infrastructures, as these constitute permanent services provided by universities and thus provide some measure of protection from the uncertainties of temporary third-party funding.

4.5 Beyond academia

The origins of language documentation are inextricably linked to the growing awareness of the magnitude of language endangerment and the threat that language loss poses to the linguistic and cultural heritage of our world. From the outset, societal aspects were an integral part of the nascent field, and were consistently articulated alongside academic interests. Societal interests were also the driving force behind the establishment of non-profit organizations that were to raise awareness, support the preservation of endangered languages, and promote language documentation activities. This interaction between academic and societal issues can be traced through the history of the *Gesellschaft für bedrohte Sprachen* [Society for Endangered Languages] (GBS; since 1997).

In 1992, the *Deutsche Gesellschaft für Sprachwissenschaft* [German Linguistic Society] (DGfS) established a project group on endangered languages, which obtained financial support from the Volkswagen Foundation and organized a summer school on *Sprachbeschreibung und Feldforschung* [language description and fieldwork] in Cologne in 1993. Many of the participants went on to play leading roles in the subsequent development of language documentation as a field. One central issue was the question of how to support local initiatives to document and preserve a community's own language. As this was outside the scope of established research funding, the idea emerged to set up a non-profit organization, funded by membership fees and donations: the DGfS project group thus became the registered association GBS, founded in 1997 in Cologne. Since then, the GBS has funded more than 90 small-scale initiatives around the world, and has undertaken activities to communicate the urgency of the problem of language endangerment and the value of linguistic diversity to the general public.

Nikolaus was one of the 21 founding members of the GBS in 1997, and he became its long-term president (2002–2008, 2010–2020). As president, he made science communication a priority, giving numerous interviews for print and broadcast media. See also Herrmann et al. (this volume) for a pilot investigation in how language documentation is represented in the media. More recently, Nikolaus

developed an academic interest in science communication — and mediated communication more generally — and teamed up with researchers from philosophy, media studies, and media psychology, as well as with practitioners from the media industry, to probe the role of concepts such as trust in public discourse (Packard & Himmelmann, in prep.).

5. Structure of this book

One of the lasting legacies of language documentation is that it has put data centre stage and advanced our theorizing about data — the contexts of data production and reproduction, the resulting data types and their characteristic properties, the insights into metalinguistic knowledge that they offer, and the kinds of research questions that they can (and cannot) be used to answer. More generally, the documentarist turn has led to an increasing awareness of the empirical foundations of our discipline, accompanied by an ever-growing appreciation of the value of observable linguistic behaviour. Today, there exist well-annotated, audiovisual data collections for many — if still far from enough — small, un(der)described languages across the world. Many of these collections strive to represent a "lasting, multipurpose record of a language" — not an end in itself, but paving the way for multiple types of uses and reuses. This means that documentary data is increasingly supporting theory-building on a large scale, which is where the present volume steps in: What does the shift towards more diverse, ecologically valid data types mean for synchronic and diachronic linguistic analyses and typological generalizations? To what extent does the careful attention to creating multimodal datasets and recording communicative events of different levels of spontaneity impact how we think about language and language structure? Does it offer us the chance to ask new research questions, or approach old questions from new angles and with new data types? Or, more succinctly — and once again in Nikolaus's words — what does it mean to really "take spoken language seriously"?

In this volume, we address these questions from four complementary perspectives.

In **Part I**, we situate language documentation within its larger context and follow up on some of the issues raised in Section 4 above: diversifying the field, archiving data, and science communication. Nikolaus was centrally involved in building bridges with Indonesia, and the first three papers focus on this region: Luh Anik Mayani and Katubi contextualize the documentarist turn within the Indonesian academic landscape (Chapter 2); Yusuf Sawaki, Hugo Warami, Apriani Arilaha, Jeanete Lekeneny, and Boas Wabia focus on the Western New Guinean context (Chapter 3); and I Wayan Arka reflects on the role of native- and non-

native-speaker linguists, and how collaborations between them and local communities have shaped knowledge production (Chapter 4). Nick Evans and Manuel Pamkal continue the theme of collaborative research from the perspective of working with legacy data in the Australian context (Chapter 5). Felix Rau addresses the pivotal role of language archives, and argues that they have evolved in parallel with developments in language documentation, responding to, and interacting with, language documentation's conceptual and methodological needs (Chapter 6). And Tobias-Alexander Herrmann, Kai Kaspar, and Stephan Packard present a pilot study investigating how language documentation is represented in science communication (Chapter 7).

Part II explores the potential of language documentation from the perspective of documentary corpora, and examines how theorizing about primary data is opening up new avenues of research. A central tenet of language documentation is to record communicative events according to different degrees of spontaneity, based on the assumption that "spontaneity is correlated with aspects of linguistic structure" (Himmelmann 1998: 180). On the one hand, this assumption brought spontaneous language use in conversational interaction to the forefront of documentation — i.e. a type of communicative event that had previously been neglected. On the other hand, this programmatic approach entails continued interest in less spontaneous forms of language use, as often found in writing but also in more monologic and ritual forms of speech and sign, as well as song language.

Starting with one end of the continuum, we note that the shift towards creating rich audiovisual records of communicative events has turned the spotlight on language in interaction, which includes illuminating the multimodal nature of language. The availability of such datasets for a wide range of languages offers the unprecedented opportunity to systematically integrate interactive and multimodal perspectives into linguistic analysis, and thus has the potential to fundamentally alter how we think about language. Mandana Seyfeddinipur, Pamela Perniss, and Claudia Wegener show what we stand to gain by taking multimodal language use seriously (Chapter 8). Anastasia Bauer, Sonja Gipper, Tobias-Alexander Herrmann, Jana Hosemann, and Simona Sbranna continue with the topic of multimodality, and address a mismatch in the corpus landscape: while documentary corpora of spontaneous spoken/signed language use are increasingly becoming available for minority languages, corpora of majority languages tend to focus on written language. Their contribution reviews developments in creating multimodal corpora, discussing how different corpus-building traditions can benefit each other (Chapter 9).

Linking to the other end of the spontaneity continuum, Silvia Kutscher shows how the careful construction of a dataset of Pharaonic Egyptian monumental inscriptions allows for the study of multimodality in historical records (Chapter 10).

Melanie Schippling reflects on the role of orality and literacy in language documentation in general, and on the importance of documenting written language and writing practices in particular (Chapter 11). Uta Reinöhl highlights the diachronic depth intrinsic to many seemingly "synchronic" documentation corpora; for example, because remotely spoken dialects may retain older forms, as do oral-tradition genres. This often untapped depth, she argues, has significant potential to diversify synchronic and especially diachronic theorizing (Chapter 12). A complementary contribution is provided by Volker Heeschen and Darja Hoenigman, who offer synchronic interpretations of song language in Papuan languages (Chapter 13). Part II ends with a contribution on the transcription process, i.e. the process of reducing the complexity of an audio/video signal into a written form. Katherine Walker, Birgit Hellwig, and Dagmar Jung present a case study on how a transcript can be used as a source of data in its own right, shedding light on the metalinguistic knowledge of speakers/signers (Chapter 14).

Parts III and IV bring together topical research on synchronic (**Part III**) and diachronic (**Part IV**) research questions with a view to exploring how the increasing availability of language documentation corpora sparks methodological advances, shapes analyses, and impacts theory building. A central theme that runs through many contributions is a re-orientation of the field towards corpus-based research: corpus-based methods are employed in the investigation of long-standing research questions, while new questions emerge that can only be studied through corpus data. An important new development in this respect is the rise of corpus-based typology, where research is conducted across different documentary collections so that typologies are based on spontaneous language use. Although corpus-based research has recently come to the fore, many contributions also highlight the continued importance of other approaches and the need for data from along the entire naturalness continuum — from elicited, to staged, to observed communicative events — to be taken into consideration. Indeed, several contributions reveal the limits of corpus-based research and/or discuss the value of more controlled and experimental data types for their research questions. Finally, a recurring theme throughout these two parts is the advance in our knowledge of individual languages and how this knowledge influences our understanding of the extent of, and limits to, linguistic diversity across all areas of research.

Part III is organized around the traditional areas of linguistic research. It starts with four chapters in the field of phonetics, phonology, and prosody. Frank Seifart presents cross-corpus research on pausing patterns in language documentation corpora and their relationship to parts of speech and default word order (Chapter 15). Constantijn Kaland explores the application of a new metric to capture macro-rhythm across typologically diverse languages, using corpora of staged and semi-natural speech (Chapter 16). The following two chapters turn

to the investigation of individual languages on the basis of different data types: Kirsten Culhane and Owen Edwards study vowel length in spontaneous Meto narratives (Chapter 17), while Janet Fletcher and Adele Gregory investigate glottal stop realization in Tahitian staged and elicited events (Chapter 18).

Part III then turns to morphosyntactic research, starting again with a cross-corpus study: Geoff Haig and Stefan Schnell investigate effects of animacy and syntactic role on referential choices across documentary corpora of different languages (Chapter 19). Following that, we present three studies on the topic of voice alternations in Austronesian languages. Drawing on the substantial increase in new cross-linguistic data, Dan Kaufman takes a fresh look at the *Ausrichtungsaffixe* [orientation affixes] to argue for fundamental differences between nominalization-derived and applicative-derived diathesis alternations (Chapter 20). Bradley McDonnell analyses variability in the realization of passive voice constructions in a documentary corpus of Besemah (Chapter 21). And Maria Bardají, Rowena Garcia, Christoph Bracks, and Evan Kidd study the acquisition of symmetrical voice systems across child language corpora of Tagalog, Indonesian, and Totoli (Chapter 22).

The final four contributions in Part III focus on various facets in the structuring of discourse. Two chapters investigate how morphosyntactic constructions support information flow in spontaneous language use: Katharina Haude studies information-structural effects that are achieved by placing a verb in an argument slot in Movima (Chapter 22), and Eva Schultze-Berndt researches the interaction between information structure and the agent-first preference in Jaminjung-Ngaliwurru (Chapter 23). Both chapters add novel analyses to long-standing linguistic debates, exemplifying how linguistics can benefit from taking well-curated documentary corpora seriously. Petra Schumacher and Klaus von Heusinger focus on the interactional dimension of discourse: they follow up on corpus-based observations that German demonstratives can be used in bridging contexts and present initial experimental evidence that they have a recognitional function in these contexts, conveying special shared knowledge between speakers and hearers (Chapter 25). Finally, Isabel Compes, Jakob Egetenmeyer, and Katja Hannß turn to the hierarchical organization of discourse and explore the use of morphosyntactic and lexical structures to segment larger discourse units and signal boundaries in Samoan and Chipaya narratives (Chapter 26).

In Part IV, we bring together contributions that link documentary data with diachronic research questions. Starting with the assumption that co-present social interaction is an important locus of language change, Sonja Gipper, Eugen Hill, and Pascal Coenen investigate conversational priming in repetitional responses within spontaneous language use and explore the predictions for historical changes that are generated by this mechanism (Chapter 27). This is followed by

two contributions that use corpus data to explore grammaticalization processes in the context of speech acts: Vanya Kapitonov and Caroline Gentens investigate the development of quotative markers in Kunbarlang (Chapter 28), and Katharina Gayler and Sonja Riesberg study medial speech verbs in Yali (Chapter 29). The final two contributions show how the enormous increase in available lexical and morphological data advance our understanding of diachronic relationships. On this basis, Antoinette Schapper investigates the origins and spread of Papuan Malay (Chapter 30), and Sander Adelaar traces the history of negators in Old Javanese and in the various social registers of modern Javanese (Chapter 31).

Steve Levinson rounds off this volume with an epilogue on *episteme* and *techne* that addresses the enormous impact of the documentarist turn on how we (can) research language.

Acknowledgements

This introduction was written with the help of a number of people. Isabel Compes's, Carmen Dawuda's, Dagmar Jung's, and Fritz Serzisko's memories of activities and developments at the Cologne department over the years found their way into Section 3. Section 4.1 is based on drafts by, and interviews with, Vera Szöllösi-Brenig and Helga Weyerts-Schweda, who shared with us their recollections of events at DobeS and the DFG, respectively. Lena Wolberg contributed her long-term experience in the management of the Cologne department (Section 4.2), and Felix Rau his deep knowledge of the Cologne archival infrastructures (Section 4.4). Many thanks to Katherine Walker whose perspective as a linguist and native speaker has greatly improved this chapter. All errors are, of course, our own.

References

Bechert, Johannes. 1990. "Universalienforschung und Ethnozentrismus." In *Proceedings of the 14th International Congress of Linguists 1987, Berlin (GDR)*, 3 vols., ed. by Werner Bahner, Joachim Schildt, and Dieter Viehweger, 2350–2352. Berlin: Akademie Verlag.

Boas, Franz. 1917. "Introductory." *International Journal of American Linguistics*, 1 (1): 1–8.

Bonmann, Svenja, Sonja Riesberg, and Nikolaus P. Himmelmann. 2024. "Differential Object Marking in Western Malayo-Polynesian Symmetrical Voice Languages." *Linguistic Typology* 29 (1): 127–165.

Bouquiaux, Luc, and Jacqueline M.C. Thomas (eds). 1976. *Enquête et description des langues à tradition orale. L'enquête de terrain et l'analyse grammaticale*, vol. 1. Paris: SELAF.

Cruz Cruz, Emiliana (ed). 2020. *Theoretical Reflections Around the Role of Fieldwork in Linguistics and Linguistic Anthropology: Contributions of Indigenous Researchers from Southern Mexico*. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24977

Evans, Nicholas, and Alan Dench. 2006. "Introduction: Catching Language." In *Catching Language. The Standing Challenge of Grammar Writing*, ed. by Felix K. Ameka, Alan Dench, and Nicholas Evans, 1–39. Berlin: De Gruyter Mouton.

Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel (eds). 2006. *Essentials of Language Documentation*. Berlin: De Gruyter Mouton.

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. 1992. "Endangered Languages." *Language* 68 (1): 1–42.

Himmelmann, Nikolaus P. 1983. Linking im Tagalog. MA thesis, LMU München.

Himmelmann, Nikolaus P. 1986. "Morphosyntactic Predication. A Functional-Operational Approach." *Arbeiten des Kölner Universalien-Projekts* 62. Köln: Institut für Sprachwissenschaft.

Himmelmann, Nikolaus P. 1987. *Morphosyntax und Morphologie. Die Ausrichtungsaffixe im Tagalog*. München: Fink.

Himmelmann, Nikolaus P. 1996. "Demonstratives in Narrative Discourse: A Taxonomy of Universal Uses." In *Studies in Anaphora*, ed. by Barbara Fox, 205–254. Amsterdam: Benjamins.

Himmelmann, Nikolaus P. 1997. *Deiktikon, Artikel, Nominalphrase: Zur Emergenz syntaktischer Struktur*. Tübingen: Niemeyer.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2000. "Towards a Typology of Typologies." *Sprachtypologie und Universalienforschung* 53 (1): 5–12.

Himmelmann, Nikolaus P. 2002a. "Voice in Two Northern Sulawesi Languages." In *The History and Typology of Western Austronesian Voice Systems*, ed. by Fay Wouk, and Malcolm Ross, 123–142. Canberra: Pacific Linguistics.

Himmelmann, Nikolaus P. 2002b. "Voice in Western Austronesian: An Update." In *The History and Typology of Western Austronesian Voice Systems*, ed. by Fay Wouk, and Malcolm Ross, 7–16. Canberra: Pacific Linguistics.

Himmelmann, Nikolaus P. 2004. "Lexicalization and Grammaticization: Opposite or Orthogonal?" In *What Makes Grammaticalization — A Look from its Components and its Fringes*, ed. by Walter Bisang, Nikolaus P. Himmelmann, and Björn Wiemer, 21–42. Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2005. "Gram, Construction, and Class Formation." In *Wortarten und Grammatikalisierung: Perspektiven in System und Erwerb*, ed. by Clemens Knobloch, and Burkhard Schaeder, 79–93. Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2006a. "Language Documentation: What Is It and What Is It Good for?" In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2006b. "The Challenges of Segmenting Spoken Language." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 253–274. Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2006c. "Prosody in Language Documentation." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 163–181. Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2008a. "Reproduction and Preservation of Linguistic Knowledge: Linguistics' Response to Language Endangerment." *Annual Review of Anthropology* 37: 337–350.

Himmelmann, Nikolaus P. 2008b. "Lexical Categories and Voice in Tagalog." In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter Austin, and Simon Musgrave, 247–293. Stanford: CSLI.

Himmelmann, Nikolaus P. 2012. "Linguistic Data Types and the Interface Between Language Documentation and Description." *Language Documentation and Conservation* 6: 187–207. http://hdl.handle.net/10125/4503

Himmelmann, Nikolaus P. 2022a. "Prosodic Phrasing and the Emergence of Phrase Structure." *Linguistics* 60 (3): 715–743.

Himmelmann, Nikolaus P. 2022b. "On the Comparability of Prosodic Categories: Why 'Stress' is Difficult." *Linguistic Typology* 27 (2): 341–361.

Himmelmann, Nikolaus P. 2022c. "Against Trivializing Language Description (and Comparison)." *Studies in Language* 46 (1): 133–160.

Himmelmann, Nikolaus P. & D. Robert Ladd. 2008. Prosodic Description: An Introduction for Fieldworkers. *Language Documentation and Conservation* 2 (2): 244–274. http://hdl.handle.net/10125/4345

Himmelmann, Nikolaus P., and Daniel Kaufman. 2020. "Prosodic Systems: Austronesia." In *The Oxford Handbook of Language Prosody*, ed. by Carlos Gussenhoven, and Aoju Chen, 370–383. Oxford: Oxford University Press.

Himmelmann, Nikolaus P., and Beatrice Primus. 2015. "Prominence Beyond Prosody — A First Approximation." In *pS-prominenceS: Prominences in Linguistics. Proceedings of the International Conference*, ed. by Amedeo De Dominicis, 38–8. Viterbo: DISUCOM Press.

Himmelmann, Nikolaus P., and Sonja Riesberg. 2013. "Symmetrical Voice and Applicative Alternations: Evidence from Totoli." *Oceanic Linguistics* 52 (2): 396–422.

Himmelmann, Nikolaus P., and Eva Schultze-Berndt. 2005. "Issues in the Syntax and Semantics of Participant-Oriented Adjuncts: An Introduction." In *Secondary Predication and Adverbial Modification*, ed. by Nikolaus P. Himmelmann, and Eva Schultze-Berndt, 1–67. Oxford: Oxford University Press.

Hockett, Charles F. 1968. *The State of the Art*. The Hague: Mouton.

Jung, Dagmar, and Nikolaus P. Himmelmann. 2011. "Retelling Data: Working on Transcription." In *Documenting Endangered Languages. Achievements and Perspectives*, ed. by Geoffrey Haig, Nicole Nau, Stefan Schnell, and Claudia Wegener, 201–220. Berlin: De Gruyter Mouton.

de Lima Silva, Wilson, and Katherine J. Riestenberg (eds). 2020. *Collaborative Approaches to the Challenges of Language Documentation and Conservation*. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24939

McDonnell, Bradley, Andrea L. Berez-Kroeker, and Gary Holton (eds). 2018. *Reflections on Language Documentation 20 Years after Himmelmann 1998*. Honolulu: University of Hawai'i Press.

Mühlhäusler, Peter. 1985. "Review of Young People's Dyirbal, by Annette Schmidt." *Linguistics* 6: 1005–1008.

Packard, Stephan, and Nikolaus P. Himmelmann (eds). In Prep. *Trust in Science Communication: A Feedback-Based Multi-Disciplinary Perspective*. (Frontiers in Communication, Research Topic.) https://www.frontiersin.org/research-topics/63742/trust-in-science-communication-a-feedback-based-multi-disciplinary-perspective

Reinöhl, Uta & Nikolaus P. Himmelmann. 2017. "Renewal: A figure of speech or a process sui generis?" *Language* 93 (2): 381–413.

Riesberg, Sonja. To Appear. "Documenting and Describing the Languages of Eastern Indonesia — Past, Present, and Future." In *Indonesian Languages and Linguistics: State of the Field*, ed. by Abigail C. Cohn, Maya R. Abtahian, J. Joseph Errington, Thomas J. Conners, and Yanti. Ithaca: Cornell University Press.

doi Riesberg, Sonja, Maria Bardají i Farré, Kurt Malcher, and Nikolaus P. Himmelmann. 2022. "Predicting Voice Choice in Symmetrical Voice Languages. All the Things That Do Not Work in Totoli." *Studies in Language* 46 (2): 453–516.

doi Riesberg, Sonja, Janina Kalbertodt, Stefan Baumann, and Nikolaus P. Himmelmann. 2020. "Using Rapid Prosody Transcription to Probe Little-Known Prosodic Systems: The Case of Papuan Malay." *Laboratory Phonology* 11 (1): 8.

doi Riesberg, Sonja, Kurt Malcher, and Nikolaus P. Himmelmann. 2019. "How Universal is Agent-First? Evidence from Symmetrical Voice Languages." *Language* 95 (3): 523–561.

Samarin, William. 1967. *Field Linguistics. A Guide to Linguistic Field Work*. New York: Holt, Rinehart and Winston.

Sandman, Sarah, Shannon Bischoff, and Jens Clegg (eds). 2022. *Voices: Perspectives from the International Year of Indigenous Languages*. Honolulu: University of Hawai'i Press. https://hdl.handle.net/10125/104355

Sasse, Hans-Jürgen. 1990. "Theory of Language Death" und "Language Decay and Contact-Induced Change: Similarities and Differences". *Arbeitspapier des Instituts für Sprachwissenschaft* 12 NF. Köln: Institut für Sprachwissenschaft.

doi Schultze-Berndt, Eva, and Nikolaus P. Himmelmann. 2004. "Depictive Secondary Predicates in Crosslinguistic Perspective." *Linguistic Typology* 8 (1): 59–131.

doi Serzisko, Fritz. 1992. *Sprechhandlungen und Pausen. Diskursorientierte Sprachbeschreibung am Beispiel des Ik*. Berlin: De Gruyter Mouton.

Stirtz, Timothy, Michael Cahill, and Philip Davison (eds). 2024. *Participatory Linguistics: Methods and Case Studies from Around the World*. Honolulu: University of Hawai'i Press. https://hdl.handle.net/10125/74780

PART I

Research context

CHAPTER 2

The development of language documentation in Indonesia

Luh Anik Mayani & Katubi
Indonesian Embassy Paris | National Research and Innovation Agency (BRIN)

This chapter highlights the development of language documentation in Indonesia, both at universities and at research institutions. It also presents the roles of linguistics associations in Indonesia as actors of supporting language documentation. Finally, it discusses what it takes to build a national language archive, and describes how a language documentation archiving project intends to preserve language diversity in Indonesia.

Keywords: language documentation in Indonesia, capacity building, teaching, archiving

1. Introduction

Language documentation, as defined by Nikolaus P. Himmelman in his seminal 1998 paper 'Documentary and Descriptive Linguistics', is a "comprehensive record of the linguistic practices characteristic of a given speech community" (Himmelmann 1998: 166). This approach also involves the use of multimedia tools, such as audio and video recordings, and it involves communities directly in documenting their languages (Austin & Sallabank 2011). It recognizes that language is closely linked to culture and, therefore, that documenting and preserving it requires more than just the written records often compiled in more traditional field-work based research. Instead, it requires active, community-based methods that capture how people use their language in everyday contexts and express their cultural identities (Grenoble & Whaley 2006). Further, the possibility to combine new technology with ethical, community-based practices serves both researchers and the language communities themselves.

The emergence of documentary linguistics is particularly important for linguistic research in Indonesia, which has remarkable linguistic diversity that is at

https://doi.org/10.1075/slcs.240.02may
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

risk of disappearing. With around 700 languages — about 430 Austronesian and 270 non-Austronesian — Indonesia is home to roughly one-tenth of all the languages in the world (Simons & Fenning 2018). Both the large number of languages and the fact that the majority of these languages are no longer transmitted to children make it urgent for local scholars to be trained and included in the endeavour of documenting and describing the languages of the region.

Probably one of the first larger workshops on language documentation methods in Indonesia was organised by Nikolaus P. Himmelmann and his team, funded through the DobeS programme of the Volkswagen Foundation. A series of two workshops, titled "*Language Documentation — Basic Training*", was held in Ubud, Bali, in 2006 and 2007. Participants were students from various universities in Indonesia and researchers from research centres, such as *Lembaga Ilmu Pengetahuan Indonesia* (LIPI, now called *Badan Riset dan Inovasi Nasional* (BRIN) — the National Research and Innovation Agency), as well as from *Badan Pengembangan dan Pembinaan Bahasa* (the Agency for Language Development and Cultivation). The two authors of this chapter also attended these events, and following the so-called 'Bali workshops', several of its alumni went abroad and successfully pursued doctoral degrees in linguistics by applying the language documentation approach, some of them under the supervision of professors who taught at the Bali workshops. Others have continued working on language documentation at their respective universities or research institutions within Indonesia.

2. Language documentation at Indonesian universities

Indonesian universities, BRIN, as well as linguistics associations have increasingly recognized the importance of language documentation for both research and pedagogical purposes. This can be seen from the organization of various workshops, seminars, and conferences focused on language documentation, often in collaboration with international partners, but also from the integration of language documentation courses in the linguistics programmes of Indonesian universities, as further elaborated in the following two sub-sections.

2.1 Workshops on language documentation at Indonesian universities in collaboration with foreign researchers

Subsequent to the Bali workshops, similar events took place in various locations in Indonesia. In this section, we list some of them in order to showcase the variety of institutions and people involved, without claiming to be exhaustive.

For example, Atma Jaya Catholic University of Indonesia hosted a number of workshops on language documentation. In March 2010 in partnership with SIL International, it conducted a workshop titled "Fundamentals of Oral Language Documentation";[1] in 2018 Atma Jaya organised a training course on ELAN and FLEx;[2] and in 2021 a webinar on Audacity and ELAN was offered with support from the American Institute for Indonesian Studies (AIFIS) and Potsdam University, Germany.[3]

Between 2012 and 2017, several workshops on general linguistics and language documentation were organised by the Center for Endangered Languages Documentation (CELD) at the University of Papua, Manokwari, in collaboration with Nikolaus Himmelmann and his team from the University of Cologne (all funded by the DoBeS programme of the Volkswagen Foundation). Furthermore, a number of teach-ins and workshops were offered by international researchers whose research was sponsored by CELD (see Sawaki et al., this volume).

Around the same time, i.e. between 2013 and 2016, a series of workshops were conducted by Asako Shiohara (Tokyo University of Foreign Studies) and colleagues as part of the LingDy programme. These workshops took place at various different Indonesian universities (Udayana, Denpasar; Mulawarman University, Samarinda; Seloko Institute, Politeknik Jambi; Manado State University; The University of Nusa Cendana, Kupang). See Jukes et al. 2017 for a full report on these events.

Similarly, Universitas Kristen Artha Wacana in Kupang, East Nusa Tenggara, organized a workshop on "Documenting Minority Languages in Nusa Tenggara Timur" in May 2018, in collaboration with the Netherlands Organization for Scientific Research, the United States National Science Foundation, and the Language Documentation Training Center at the University of Hawaii.[4] See Jacob, Klamer, and Holton (2018) for the outcomes of this event.

Additionally, some workshops on language documentation were conducted in partnership with AIFIS. Among others, in 2020 a workshop was held by a collaborative team from AIFIS, PKBB-Atma Jaya, Universitas Brawijaya, and Nick Williams (Colorado University Boulder).[5] In August 2021 with support from AIFIS, Universitas Brawijaya in Surabaya hosted a webinar on "Language Documentation

1. https://www.sil.org/about/news/language-documentation-workshop-indonesia
2. https://www.atmajaya.ac.id/id/media/BuletinPKBB_Semester_II_2018.pdf
3. https://www.atmajaya.ac.id/id/pages/2521-elan-jan/
4. https://scholarspace.manoa.hawaii.edu/items/82a6c60d-6f93-408d-8a6d-0c5b5d05f9d2
5. https://www.aifis.org/lectures-and-workshops/2020/9/22/online-workshop-on-elan-for-language-preservation

and Studies of English Spoken in Indonesia".[6] In October 2021, with the assistance of AIFIS and funded by a grant from the Documenting Endangered Language Program of the National Science Foundation awarded to the University of Delaware, a linguistic fieldwork and documentation training event took place in Kupang, Indonesia, where eight Indonesian participants and four US graduate students formed teams and were trained to collaborate in documenting and describing local languages spoken in one of the communities found on West Timor and its surrounding islands. This project was also run in collaboration with the Atma Jaya Catholic University of Indonesia as well as the Research Institute for Languages and Cultures of Asia and Africa (ILCAA), an affiliate of the Tokyo University of Foreign Studies (TUFS).[7]

Recently, in January 2025, Udayana University in Bali conducted a workshop on "Language Documentation Methods through Elan and FLEx" as part of the Enggano language documentation project. The project team was represented by Charlotte Hemmings from the Faculty of Linguistics, Philology, and Phonetics, University of Oxford.[8]

2.2 Teaching language documentation in linguistics programmes at Indonesian universities

The teaching efforts listed in the previous section were one-off events that partly targeted university students, but often also local researchers and community-internal language activists. In recent years, a more sustainable developmental pathway can be observed, namely the integration of language documentation in academic teaching within universities' linguistics programmes, in some instances even as a subject of its own right integrated into university programmes' curricula. We consider this a positive trend, even though the number of institutions that include language documentation as a permanent course in their curricula can certainly still be increased.

As an institutionalized programme, i.e. integrated into university curricula, it seems that the University of Papua in Manokwari was the first among Indonesian universities to offer courses in language documentation and preservation. The course began in 2008 as a 3-credit module for undergraduate students in

6. https://www.aifis.org/lectures-and-workshops/2021/7/29/language-documentation-and-studies-of-english-spoken-in-indonesia

7. https://sites.udel.edu/fieldwork-training-indonesia/

8. https://fib.unud.ac.id/posts/bachelor-of-english-literature-holds-workshop-on-language-documentation-methods-through-elan-and-flex-as-part-of-two-days-of-the-enggano-language-documentation-and-lexicography-a-reflection

the English programme. This course combines theoretical knowledge and practices, providing students with opportunities to document narratives or even simple topics, such as body parts in indigenous languages, through video recording. These videos are uploaded to a YouTube channel, featuring subtitles not only in the original language, but also in Indonesian and English. Such practices enable the documentation of indigenous languages in Papua and encourage students to be creative in selecting topics and creating videos (p.c. Yusuf Sawaki; lecturer in the English programme, University of Papua).

The University of Jambi, too, is among the few institutions offering a dedicated course for undergraduate students — specifically, a 3-credit module on language documentation within the Indonesian literature programme.[9]

In addition, based on personal communication with several colleagues, Universitas Indonesia in Jakarta offers a 3-credit course on language documentation and revitalization within its Master of Linguistics programme (p.c. Nazarudin; lecturer at the Linguistics Department, Faculty of Humanities, Universitas Indonesia). In East Nusa Tenggara, the same course is offered as a 2-credit module in the Master of Linguistics programme, Universitas Nusa Cendana, as well as in the undergraduate English Education programme, Universitas Kristen Artha Wacana (p.c. Jermy Balukh; lecturer in the Master of Linguistics programme, Nusa Cendana University).

Other universities tend to offer language documentation as a topic within more general courses. For example, at Atma Jaya Catholic University of Indonesia, it is introduced within the course "Current Issues in Linguistics" in the master's programme in Applied English Linguistics. Although presented as part of a broader curriculum, 4–5 sessions are dedicated to demonstrating how the language documentation approach is applied to compile and process linguistics data, including practical activities (collecting data, ELAN, data processing, and basic analysis). These sessions are essential for master's students to understand language documentation as an approach that emphasizes the use of multimedia tools, such as audio and video recordings, followed by annotation, a crucial step toward linguistic analysis (p.c. Yanti; lecturer in the Applied English Linguistics programme, Atma Jaya Catholic University of Indonesia).

Moreover, Udayana University in Bali introduces language documentation starting from the undergraduate level within the English Literature programme as part of the research methodology course. Furthermore, at the postgraduate level, including master's and doctoral programmes within the linguistics programme, language documentation is integrated into the sociolinguistics course, where stu-

9. https://sastraindonesia.unja.ac.id/index.php/kurikulum-sastra-indonesia/

dents are trained to apply the language documentation approach in their sociolinguistic research (p.c. Ni Nyoman Seri Malini; lecturer in the English Literature programme, Udayana University).

To conclude, from the initial efforts of teaching language documentation in Indonesia through pioneering events like the "Bali workshops", to the various subsequent workshops, seminars, and collaborations with international institutions, awareness and academic engagement with language documentation have grown. Today, several universities and research institutions conduct training and research in this field. However, the integration of language documentation into university curricula remains limited, with the University of Papua in Manokwari standing out as a pioneer, offering a dedicated, practical course since 2008. While some universities integrate language documentation topics into broader courses, there is still a need for more widespread, institutionalized programmes to ensure the sustainable preservation of Indonesian linguistic diversity.

3. Language documentation at research and language institutions

3.1 The national research and innovation agency (BRIN)

Also at the research level, documenting Indonesia's many languages is a key enterprise. While for a long time many institutions have focused on sociolinguistic surveys and the collection of word lists, in recent years a shift towards modern language documentation and the collection of multi-media language corpora has been taking place. The National Research and Innovation Agency (BRIN) supports these efforts with dedicated facilities and programmes. In particular, its Research Center for Language and Literature Preservation, established in 2022, has declared the documentation of ethnic languages categorized as endangered to be one of its major goals.

As many BRIN researchers previously had only little experience with language documentation methods, the Research Center for Language and Literature Preservation — BRIN conducted language documentation training, in collaboration with, among others, Tokyo University for Foreign Studies (TUFS), Japan, through Asako Shiohara, and the University of Cologne, Germany, through Nikolaus Himmelmann, Sonja Riesberg, and colleagues. Starting from the basic concepts of language documentation, the training covered special language-documentation software and the practice of language documentation.

Currently, the following languages are being documented by the Research Center for Language and Literature Preservation — BRIN:

Saparua:	1,500 speakers, Saparua island, Central Maluku
Alune:	10,000 speakers, Seram, Central Maluku
Wabo:	100 speakers, Yapen, Papua
Modole:	3,000 speakers, Halmahera, North Maluku
Modo:	2,000 speakers, Labuan Bajo, East Nusa Tenggara
Kui:	800 speakers, Alor island, East Nusa Tenggara
Dayak Tunjung:	10,000 speakers, East Kalimantan
Dayak Bentian:	4,500 speakers, East Kalimantan
Limola:	100 speakers, Suwuk, South Sulawesi
Waili:	500 speakers, Raja Ampat, Papua
Koiwai:	2,000 speakers, Kaimana, West Papua
Mairasi:	4,000 speakers, Kaimana, West Papua
Kamberaw:	No data available, Kaimana, West Papua
Madewana:	No data available, Kaimana, West Papua
Wawoni:	No data available, Wawoni Island, Konawe, Southeast Sulawesi

In addition, while researchers from the Research Center for Language and Literature Preservation — BRIN work in the field to document endangered languages, they also provide language documentation training to members of the respective language communities. The benefits of successfully finding and training community language activists in language documentation are numerous:

1. The documentation (linguistic and otherwise) is richer. This is due to the speed with which language community members can document their own language compared to outside researchers. E.g., community members possess native-speaker meta-linguistic knowledge when transcribing and translating language documentation recordings.
2. Indigenous community awareness can be strengthened. Even if the community's interest in their local language is initially low, after working together on the documentation of the language, their awareness of preserving their own language can grow stronger.
3. Language revitalization and its outcomes are better understood. By discussing various possibilities for utilizing the results of language documentation to revitalize their own languages, the usefulness of this approach is better understood.
4. Sustaining language work based on the energy of grassroots community support. This is crucial because grassroots support from language communities is the foundation of successful language documentation and long-term language revitalization.

5. Training increases the awareness of linguistic diversity and can reduce the gap between linguists and community members. A close relationship between researchers and members of the language community helps in many ways. A better understanding of the benefits of linguistic diversity can, for example, lead to stronger engagement from community members in safeguarding their mother tongue.

Based on these observations, there are a number of strategies that need to be implemented to develop language documentation in Indonesia. First, human resource capacity needs to be increased. Training and education are needed to boost the number of well-trained linguists from Indonesia in language documentation. Second, language communities need to be involved more actively in the language documentation process. And finally, the regional governments need to become aware of their responsibility to protect local languages as mandated by law, and take appropriate measures in this regard.

Besides the human resource challenges for language documentation, Indonesia also faces issues related to archiving. Indonesia currently lacks a centralized language documentation archiving centre. Before 2023, Indonesia did not have a digital repository for preserving multimedia collections resulting from the documentation of endangered languages. Given the extraordinarily high number of indigenous (endangered) languages (see Section 1), and the increasing efforts in the field of modern language documentation (see above), such an archive is, however, urgently needed. In Section 4, we will discuss in detail what it takes to build a national language archive for Indonesia.

3.2 The language development and Cultivation Agency

Another key player in language documentation and revitalization in Indonesia is the Language Development and Cultivation Agency (*Badan Pengembangan dan Pembinaan Bahasa* or shortened to *Badan Bahasa*), a governmental agency under the Ministry of Primary and Secondary Education of the Republic of Indonesia. It is responsible for supporting the development and promotion of the Indonesian language as well as the strengthening and revitalization of local indigenous languages.

Its initial focus was on conducting research on language mapping, started in 1991–1997. After a prolonged pause due to the monetary crisis, the project was resumed in 2006 and continued until 2019 with up to 2,560 research locations across Indonesia. The first edition of Badan Bahasa's language map was published

in 2008 and the latest (sixth) edition was released in 2019. According to this last edition, Badan Bahasa mapped 718 local languages in Indonesia.[10]

In addition to the language map, Badan Bahasa also published local language dictionaries, sketch grammars, and other linguistic resources, which can be considered as complementary results of nearly two decades of language mapping activities. On a practical level, the agency also works on creating children's short stories in local languages, which aim to preserve both the language and the traditional local knowledge.[11]

Furthermore, being within the Ministry of Primary and Secondary Education enables Badan Bahasa to develop language materials for indigenous languages, e.g. children's short stories, and implement them into school curricula. They thereby play an important role in ensuring that Indonesia's linguistic diversity is valued and preserved.

3.3 The role of linguistics associations

The Indonesian Linguistics Society (or *Masyarakat Linguistik Indonesia*, MLI), founded in 1975, is arguably the oldest linguistics society in Indonesia. Its membership consists of researchers, professors, linguists, lecturers, students, observers, and anyone interested in linguistics studies, not only from Indonesia but also from abroad. MLI consistently supports the development of linguistics studies in Indonesia through various activities, including organizing conferences, workshops, and seminars that focus on new methodologies and current issues in the field.

In the area of language documentation, MLI promotes collaboration among linguists, society members, and other stakeholders, raising awareness and building capacity among researchers to preserve endangered languages. In November 2018 in collaboration with the Language Development and Cultivation Agency and the University of Cologne (the latter represented by Nikolaus Himmelmann, Maria Bardají, and Christoph Bracks), MLI hosted a workshop on language documentation. This event was attended not only by researchers from the agency, but also by researchers from LIPI (now BRIN) and independent researchers seeking to update their skills and knowledge concerning language documentation theories and practices. Although MLI does not have specific projects on language documentation, it continues to support the field through webinars and seminars.

---

10. The map is accessible online at https://petabahasa.kemdikbud.go.id/index.php. This website provides information on the area where each language is spoken including its dialects, if available.

11. An example of these materials is available at this link: https://budi.kemendikdasmen.go.id/bacabuku/digital/zamrud-khatulistiwa#your_book_name/1.

In 2020, during the Covid-19 pandemic, the society launched the "Online Linguistics Class", which featured presentations by linguists or other researchers from Indonesia and abroad on recent research, trends, and approaches in linguistics. One of these classes focused on language documentation and revitalization (presented by Asako Shiohara from Tokyo University of Foreign Studies and Yanti from Atma Jaya Catholic University of Indonesia), while another discussed annotation methods and discourse studies based on language documentation corpora (presented by Maria Bardají (University of Barcelona), Katherine Walker (University of Cologne), and Sonja Riesberg (CNRS and University of Cologne).

Additionally, at its biennial international linguistics conference KIMLI, MLI consistently invites speakers to share their research on language documentation or revitalization. Moreover, in its linguistics journal *Journal Linguistik Indonesia*, one of the topics within its scope is language documentation.

4. Creating a national language archive for Indonesia

In 2023, the Research Center for Language and Literature Preservation — BRIN in collaboration with the Directorate of Repositories, Multimedia and Scientific Publications (RMPI)-BRIN initiated the design of a Repository Center for Language Documentation in Indonesia, called LADIN (Language Documentation in Indonesia). LADIN is a digital repository tool developed by BRIN that facilitates the long-term storage of language documentation corpora compiled through research in Indonesia, as well as access to this data.

LADIN is the first national language archive in Indonesia, and given that it is still in its infancy, LADIN faces a number of challenges. First, the technical staff working on the repository, while being experienced with general research archives, have little familiarity with designing an archive for spoken language. However, archiving indigenous languages requires specialized skills and technical requirements, as further discussed below. Hence, training is needed for those building and using this archive. We therefore formed a team consisting of archive staff (Repository, Multimedia, and Scientific Publication — BRIN) and researchers at the Research Center for Language and Literature Preservation — BRIN. The following steps will be undertaken in the near future:

1. Training will be conducted once LADIN is ready to launch.
2. The team will help researchers at BRIN, language community members, and the interested public to use LADIN.
3. Exploring the possibility of collaborating with well-established archives such as the Kaipuleohone Language Archive (University of Hawai'i), ELAR, etc.

4. Foreign researchers will be informed that audio and video recordings are considered material subject to the Material Transfer Agreement (MTA) regulations, and that archiving with LADIN will be requested at some point in the future.

In order for LADIN to properly develop as a centre for archiving language documentation materials in Indonesia, there will be certain obligations for all researchers, both Indonesian and foreign. Indonesian and foreign researchers will be required to store their language documentation results in LADIN, managed by the LADIN team/BRIN. The importance of mandatory submission and mandatory storage (WAJIB SERAH WAJIB SIMPAN) of this data is emphasized in the National Science and Technology System Law (article 40). Data and research outputs must be stored for at least 20 years. Further arrangements are mandated in specific government regulations (Peraturan Pemerintah).

There are three principles underlying the creation and maintenance of the LADIN repository: First, such an archive needs to protect both the languages and the speech communities, not only knowledge and intellectual property. Second, an archive requires continuous upkeep and updates. Third, a national archive must follow Indonesian laws and regulations, which need to be carefully applied to LADIN. These three points will be briefly discussed one by one below.

First, the archive needs to protect both the languages and the speech communities, not only knowledge and intellectual property. This is done by incorporating ethical, participatory, and reparative practices. This approach moves away from the traditional, extractive model of archiving towards one that empowers communities and supports language revitalization. Ethical archiving recognizes that data from a language is inextricably linked to the community that speaks it. It requires archivists to prioritize the rights, values, and well-being of the speech community at every stage of the process. Key ethical practices include prior informed consent, data sovereignty and governance, repatriation and restitution, and the right of reply:

- Prior informed consent: Ethical archives go beyond standard copyright. They secure free, prior, and informed consent from the community, especially when working with indigenous languages. This recognizes that cultural information may be confidential, secret, or subject to customary laws that supersede standard intellectual property rights.
- Data sovereignty and governance: Indigenous communities have the right to maintain and control their cultural heritage and intellectual property. Archives must facilitate this right, allowing communities to make decisions about how their language materials are accessed, used, and stored. This can

involve giving communities the option to restrict certain content or define culturally appropriate access protocols.
- Repatriation and restitution: For archives containing materials collected historically without community consent, archivists have a responsibility to facilitate the return of the materials to the rightful owners. This process is crucial for healing historical trauma and rebuilding trust.
- Right of reply: The Indigenous Archives Collective asserts the rights of indigenous peoples to challenge and respond to the information held about them in archives. Archives can protect communities by creating frameworks for a "right of reply" that addresses potentially harmful or inaccurate historical records.

LADIN–BRIN, as a centre for archiving language documentation data in Indonesia, provides support for languages through community-based revitalization, because effective language archives must move beyond passive storage to actively supporting language revitalization efforts. This involves working with communities to create resources that keep the language vibrant and in use. Key language revitalization practices include participatory archiving, co-designing learning materials, promoting inter-generational transmission, and creating accessible resources.

- Participatory archiving: This approach shifts power by involving community members as co-creators of the archive, rather than just subjects. This ensures that the collection reflects the community's priorities and needs.
- Co-designing learning materials: Archival recordings and linguistic data should be used to create educational materials, tools, and curricula that support language learners and teachers. This includes resources for listening, speaking, reading, and writing, tailored to the community's learning goals.
- Promoting inter-generational transmission: By making archived materials accessible and useful, archives can encourage elders and other fluent speakers to continue using and passing on the language to younger generations.
- Creating accessible resources: Archives can support language health by providing practical resources like phrasebooks, dictionaries, and recordings. Critically, these materials must be made available in ways that the community can easily access, such as through websites, apps, or in local cultural centres.

Because language is more than a collection of words and grammar, LADIN also aims to be a repository of cultural knowledge, history, and identity. LADIN must protect these elements in the following way: documenting contextual knowledge, preserving traditional knowledge, and countering historical misrepresentation.

- Documenting contextual knowledge: Archiving should include not only linguistic data but also the stories, histories, and cultural protocols associated with the language. This includes traditional stories, songs, and customary laws that are embedded within the language itself.
- Preserving traditional knowledge: Language often holds invaluable knowledge about the natural environment, flora, and fauna, as well as traditional medicine and sciences. By protecting the language, archives help safeguard this specialized and often irreplaceable traditional knowledge.
- Countering historical misrepresentation: Archives can work with communities to correct past inaccuracies and misrepresentations of their culture and language. By incorporating community perspectives and a "right of reply", archives ensure that the historical record is accurate and respectful.

Second, to function effectively, LADIN–BRIN as a centre for archiving language documentation data requires continuous upkeep and updates. This process involves addressing technological changes, enriching collections with new data, and engaging in ongoing ethical collaboration with speech communities to remain relevant and useful for language revitalization.

Digital and physical materials are susceptible to decay and obsolescence, demanding proactive management to ensure their longevity. To overcome this, LADIN needs the following kinds of technical maintenance: data migration, refreshing media, integrity checks, and legacy material.

- Data migration: As file formats and hardware become obsolete, data must be regularly migrated to modern, stable formats to prevent corruption and ensure access. For instance, an archive may migrate older proprietary software files to open, standardized formats like PDF/A.
- Refreshing media: Physical and digital storage media have a finite lifespan. Archives must continually refresh their holdings by transferring data to new media to avoid data loss.
- Integrity checks: Regular checksums are needed to verify that digital objects have not been corrupted during storage or transfer, ensuring the integrity of the data over time.
- Legacy material: Archives must also invest in digitizing and re-contextualizing older analogue materials, such as tape recordings and field notes, to make them accessible to contemporary communities and researchers.

Archiving is not just a technical process but a long-term, reciprocal relationship with the speech community, which is necessarily based on reciprocity, evolving needs, community governance, and addressing legal and ethical issues.

- Reciprocity: Archives must actively engage in reciprocal partnerships by not only taking data but also providing resources back to the community for revitalization efforts, such as language learning tools and curriculum development.
- Evolving needs: As communities change, their needs from the archive will also change. A passive archive becomes irrelevant; an active one adapts to support new revitalization initiatives; for instance, by creating immersive VR experiences for learners.
- Community governance: Long-term sustainability relies on community governance. This ensures that the archive's policies on access, preservation, and use evolve according to the wishes of the community itself, not just the archives' own institutional goals.
- Addressing legal and ethical issues: As new records are created and technology evolves, new legal and ethical considerations arise. An archive must continuously review and update its protocols regarding consent and intellectual property to protect the rights of community members in the digital age.

Third, for LADIN–BRIN, carefully applying Indonesian laws and regulations requires navigating a complex legal framework encompassing general archives, cultural heritage, intellectual property, and institutional research policies. The archive must balance national archiving responsibilities with the unique cultural and communal rights associated with documenting indigenous languages.

In the following, we will give a brief overview of the key Indonesian laws applicable to LADIN–BRIN and the specific considerations for their application:

1. Law No. 43 of 2009 on Archiving (*Kearsipan*)
   This law outlines the general principles and responsibilities for managing national archives, primarily under the authority of the National Archives of the Republic of Indonesia (ANRI).
   - National responsibility: The law mandates the state to organize and preserve archives as the nation's collective memory. As a state-affiliated body, LADIN's activities fall under this mandate, legitimizing its documentation efforts.
   - Archival principles: LADIN must adhere to principles of authenticity, integrity, and accessibility. However, applying "accessibility" to indigenous language materials requires a nuanced, community-centric approach that may restrict access to certain items, balancing national interest with community sensitivities.
   - Data transfer and preservation: The law establishes rules for transferring archival records. LADIN must follow these rules when acquiring and preserving language documentation collected by researchers and institutions, including materials that may have previously been held abroad.

2. Law No. 5 of 2017 on the Advancement of Culture (*Undang-undang Pemajuan Kebudayaan*)
   This law provides a formal legal basis for Indonesia's cultural advancement policy, emphasizing the preservation and development of national culture.
   - Cultural advancement: LADIN's work directly contributes to this law's goal of advancing Indonesian culture by documenting languages, which are recognized as vital national cultural assets.
   - Safeguarding indigenous culture: The law mandates that the state ensures the freedom of society to preserve and develop cultural values. This constitutional commitment under Article 32 of the Indonesian Constitution creates a legal obligation for LADIN to protect and support indigenous linguistic communities.

3. Law No. 28 of 2014 on Copyright (*Hak Cipta*) and Government Regulation No. 56 of 2022 on Communal Intellectual Property (*Kekayaan Intelektual Komunal*)
   This framework is crucial for distinguishing and protecting individual vs communal intellectual property rights.
   - Individual copyright vs Traditional Cultural Expressions (TCE): The Copyright Law primarily protects individual works for a limited time. However, Article 38 recognizes that the state holds the copyright for traditional cultural expressions (Ekspresi Budaya Tradisional).
   - Communal Intellectual Property (CIP): Government Regulation No. 56 of 2022 further establishes a framework for protecting Communal Intellectual Property (KIK). LADIN must apply this regulation by:
   - Honouring communal rights: Archiving protocols must respect the communal nature of language, stories, and oral traditions. This means not treating communal knowledge as individual intellectual property.
   - Involving communities: Archiving policies must align with the regulation's intent to encourage registration and protection of indigenous knowledge, with the communities' involvement and consent.
   - Ethical access and use: LADIN must ensure that access and use of archived materials honour the values of the communities they originate from, particularly concerning sensitive or sacred traditional knowledge.

Applying the legal framework to LADIN–BRIN, two things are of crucial importance. First, ethical acquisition and community protocols, which consist of community consent and customary laws, have to be applied. That is, any depositor in LADIN must obtain free, prior, and informed consent (FPIC) from the speech community before documenting their language. This goes beyond simple permission and requires a transparent process that explains the archive's purpose and the

community's rights. The archive should also incorporate indigenous customary laws (*adat*) into its protocols, particularly regarding access and use of culturally sensitive material. This requires direct, long-term collaboration with community elders and representatives.

Second, the archive's metadata practices in LADIN–BRIN must be culturally sensitive, using descriptions that align with community self-representation rather than imposing external classifications. Language documentation data archived in LADIN–BRIN should also include information on appropriate use and access restrictions.

By carefully implementing these legal principles in a way that respects the specific needs and rights of indigenous communities, LADIN–BRIN can ensure its repository truly serves as a protective and revitalizing resource, rather than merely a passive collection of linguistic data.

5. Summary

This chapter has given a brief overview of the state of the art of language documentation at Indonesian universities, research agencies, and linguistic societies, and demonstrated the development of the field since the first training events in documentary linguistics took place at the beginning of this century. Of course, there are still hurdles — like limited funding, technology, and trained personnel — that can slow down these efforts. But the increasing availability of digital tools and a global and national interest in multimedia language documentation open up new opportunities for expanding and improving these projects.

Nikolaus Himmelmann has been contributing to this development for almost twenty years by regularly providing workshops and training, by supervising Indonesian PhD students, and engaging in capacity-building measures for Indonesian students and community members (see also Sawaki et al., this volume).

Overall, documentary linguistics is becoming an important part of university teaching at linguistics departments and in linguistic research in Indonesia. With support from groups like MLI and the government, these efforts are helping to protect the country's rich linguistic heritage. By providing practical skills, encouraging community involvement, and promoting thorough documentation, Indonesia's universities and agencies are playing a vital role in safeguarding its many languages for the future.

Acknowledgements

Terima kasih kepada Prof. Nikolaus P. Himmelmann yang telah berjasa dalam pengembangan dokumentasi bahasa di Indonesia. Secara khusus, kami sangat berterima kasih atas inisiatif dan kontribusi Bapak melalui berbagai Workshop on language documentation, yang pertama kali diadakan di Bali pada 2006 dan dilanjutkan pada 2007. Setelah itu, Prof. Nikolaus P. Himmelmann membuat Workshop on language documentation di Badan Bahasa, di berbagai kampus di Indonesia, dan juga di BRIN.

Workshop on language documentation tersebut tidak hanya mentransfer pengetahuan teoretis, tetapi juga memberikan keterampilan praktis yang sangat berharga bagi para peneliti, dosen, dan mahasiswa di Indonesia. Metode dan bimbingan langsung dari Prof. Nikolaus P. Himmelmann dan tim telah memperkuat kapasitas kami dalam mendokumentasikan bahasa-bahasa di Indonesia secara metodis dan sistematis.

Dampak nyata dari workshop tersebut terasa melalui kualitas dokumentasi bahasa yang lebih baik dan lahirnya generasi baru ahli bahasa yang terinspirasi oleh Prof. Nikolaus P. Himmelmann. Kami sangat menghargai jasa Prof. Nikolaus P. Himmelmann dalam membangun kapasitas sumber daya manusia di Indonesia dalam bidang dokumentasi bahasa.

References

Austin, Peter K., and Julia Sallabank. 2011. "Introduction." In *The Cambridge Handbook of Endangered Languages*, ed. by Peter K. Austin, and Julia Sallabank, 1–24. Cambridge University Press.

Grenoble, Lenore A., and Lindsay J. Whaley. 2006. *Saving Languages: An Introduction to Language Revitalization*. Cambridge University Press.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics*, 36 (1): 161–195.

Jacob, June, Marian Klamer, and Gary Holton (eds). 2018. *Workshop on Documenting Minority Languages in Nusa Tenggara Timur Indonesia*. Kupang: UKAW Kupang. http://hdl.handle.net/10125/58880

Jukes, Anthony, Asako Shiohara, and Yanti. 2017. "Collaborative Project for Documenting Minority Languages in Indonesia and Malaysia." *Asian and African Languages and Linguistics* 11: 45–56.

Simons, Gary F., and Charles D. Fennig. 2018. *Ethnologue: Languages of the World*. 21st edn. Dallas, Texas: SIL International.

CHAPTER 3

Collaborative research

Language documentation and the documentarist turn in Western New Guinea

Yusuf Sawaki, Hugo Warami, Apriani Arilaha,
Jeanete Lekeneny & Boas Wabia
CELD, Universitas Papua

With its nearly 300 languages — around 219 non-Austronesian and 57 Austronesian — Western New Guinea is home to about 38 percent of all languages in Indonesia (Lewis et al. 2014). Against the background of this immense linguistic diversity, this chapter investigates how the documentarist turn has entered the Indonesian research landscape, especially in Western New Guinea (WNG). This article first provides an overview of the linguistic and sociolinguistic landscapes in WNG, highlighting the importance of language documentation in the region. Second, it discusses different efforts on the part of various institutions to document languages in WNG using classical or traditional methods. Third, it discusses the era when modern language documentation was introduced to the Indonesian and Western New Guinea research landscape, which then contributed to the establishment of the Center for Endangered Languages Documentation (CELD), the country's first centre for the documentation of indigenous languages in Indonesia.

Keywords: CELD, capacity building, language documentation, Western New Guinea

1. Introduction

This chapter discusses concepts and practical activities relating to language documentation in Western New Guinea (WNG). The region is politically part of the Republic of Indonesia and is known as West Papua or Tanah Papua. In this chapter, it is simply referred to as Western New Guinea (WNG). This geographical term encompasses six provinces in WNG: Papua Province, West Papua, Highland Papua, Central Papua, South Papua, and Southwest Papua.

https://doi.org/10.1075/slcs.240.03saw
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company



Map 1. Map of Western New Guinea, also known as West Papua or Tanah Papua, and its provincial boundaries

Western New Guinea is ethnolinguistically a region of very high linguistic diversity. There are nearly 300 languages, divided into two major language groups: the Austronesian family and the Papuan languages.[1] The Papuan languages dominate in terms of the number of languages and their distribution. This region has had a very long history of human occupation and inter-group contact that is predicted to have started approximately 50,000 years ago (Bellwood 1997; Lilley 1992; Spriggs 1996; Ricaut et al. 2024: 49). Although Foley (2000: 358) mentions that greater contact among human groups might be expected to impede, rather than foster, linguistic diversification, some areas, such as in the Cenderawasih Bay or Yapen Island, show the reverse dynamic, in that the long history of social and lin-

1. Here and below we refer to 'Austronesian' and 'Papuan' as the two major language groups in the area. It should be kept in mind, however, that only the Austronesian languages constitute a language family, while the Papuan languages are a very heterogeneous group that subsumes various different language families.

guistic contact among Papuan communities, who tend to live in small, egalitarian, clan-based societies, is probably the reason for high level of linguistic diversification (Sawaki & Arka, to appear; Foley 1986, 2000; Stasch 2013; Holton & Klamer 2018; Evans et al. 2018; Sawaki 2016a; Blust 2024: 35–48; Kamholz 2024: 181–187).

This diverse ethnolinguistic condition is also reflected in the many social, cultural, and political dynamics at the local level, which have consequences on the sociolinguistic situation today. With demographic changes, modernization, access to transportation, and advances in technology and communication, the WNG region is increasingly open. Consequently, many language groups are experiencing a shift in vitality from strong to moderate or even weak. This sociolinguistic phenomenon will be explained further in Section 2.

This chapter is organized as follows. In Section 2, we give an overview of the current linguistic landscape in WNG, comparing the facts of sociolinguistic profiles and the current sociolinguistic situation. We then provide a brief history of language description and language documentation in WNG, giving tribute to the linguistic work that has been done in this region in Section 3. In Section 4, we discuss modern language documentation in WNG and Nikolaus Himmelmann's influence on language documentation in WNG. One of his contributions is the establishment of the Center for Endangered Languages Documentation in Manokwari as a strategy for building sustainable language documentation in Western New Guinea, which is described in Section 5. In Section 6, the future challenges of CELD are articulated. In Section 7, we conclude the chapter with some final remarks, stressing how crucial language documentation is for the linguistic situation and sociolinguistic future in WNG.

2. Overview of the linguistic and sociolinguistic landscapes in WNG

This section discusses the linguistic and sociolinguistic landscapes in WNG by presenting the linguistic facts and current sociolinguistic conditions in WNG. This region is home to approximately 276 languages (Lewis et al. 2014) out of a total of approximately 850 to 1,200 languages on the island of New Guinea, comprising West New Guinea and the independent state of Papua New Guinea (PNG) (Ross 2017: 758). The languages in WNG account for about 33% of languages on the island of New Guinea, with Papuan languages dominating over Austronesian ones, which are mostly distributed in the islands and coastal areas, particularly on the north coast of WNG, around Cenderawasih Bay, and to the west around the Raja Ampat islands and Bomberai peninsula (Lewis et al. 2014; Gasser, Arnold & Kamholz 2024; Kamholz 2024) as shown in Table 1.

Table 1. Number of languages in WNG by language group

No	Language groups	Number of languages	%
1	Papuan	219	79
2	Austronesian	57	21
Total		**276**	**100**

Papuan languages are divided into several unrelated families covering about 219 languages, distributed from the Bird's Head region to the west to the area around the border of Papua New Guinea to the east. Austronesian languages fall into three groups, i.e., South Halmahera-West New Guinea (SHWNG), Bomberai, and Oceanic, as illustrated in Table 2.

Table 2. Distribution of languages in WNG by language family

No	Group	Family/sub-family	Number of langs.	%
1	Papuan	Sepik-Ramu	2	79%
		Sko	1	
		West Papuan, Bird's Head	16	
		Lakes Plain & Geelvink Bay	37	
		Trans New Guinea	154	
		Total Papuan	**219**	
2	Austronesian	SHWNG	34	21%
		Oceanic (Sarmi & Jayapura Bay)	13	
		CMP: Bomberai (north & south)	10	
		Total Austronesian	**57**	
Total languages of WNG			**276**	**100%**

Out of a total of 726 languages in Indonesia, the number of languages in Papua reaches 38% (Lewis et al. 2014; Kluge 2014: 1).

In addition to the diversity of language families and individual languages, the languages in WNG are also highly diverse at the linguistic structure level. Austronesian languages generally have a relatively simple morphological system, particularly in their verb and noun paradigms. The typical verb morphology system is a subject marker attached to the verb (Silzer 1983; van den Heuvel 2006; Mofu 2008; Gasser 2014; Sawaki 2016a; Arnold 2018). As Papuan languages comprise different language families, it is not surprising that they are more diverse in terms of their linguistic features. There are isolating languages such as Abun, Konda

(Yaben), and Yahadian in the Bird's Head region of WNG (Berry 1995; Sawaki 2018; de Vries 2004; Holton & Klamer 2018) that coexist with morphologically simple to complex languages, especially at the verb paradigm level. Some Papuan languages that have a very simple morphology are languages of the West Papuan family, found in the Bird's Head (Flassy 1991; Dol 2007; Gravelle 2010a; 2010b; Ode 1997; Reesink 1999, 2002). Other Papuan languages, especially from the Trans New Guinea family, such as Mairasi (Peckham 1982), Ekari (Mee) (Doble 1987), Western Dani (Barclay 2008), Lower Grand Valley Dani (Bromley 1981), Sentani (Cowan 1965; Hartzler 1976), Asmat (Drabbe 1959), Marind (Olsson 2021; Ndiken 2011), and others, have complex verb systems.

There are also languages with tonal systems, such as the Abawiri languages of the Lakes Plain group (Yoder 2018) and the Skou language family (McPherson & Dryer 2021). Some Trans New Guinea languages (Donohue 2005) have tone, as do a few Austronesian languages (Gasser, Arnold & Kamholz 2024), but there are also many languages in both groups without any lexical prosodic structure. Languages with voice alternation — active vs. passive contrast, like Papuan Malay[2] (Kluge 2014; Karubaba 2017) — coexist with languages without voice alternation, as in Lower Grand Valley Dani (LGVD) (Bromley 1981) and Wano (Burung 2017), representing Papuan languages, and Wooi (Sawaki 2016a, 2018) and Ambai (Silzer 1983) representing Austronesian languages. Some languages have flexible argument structure properties (Kluge 2014), and coexist with languages that have fixed argument structure (Sawaki 2016a). The diversity of languages in WNG is thus not just represented by the number of different language families and individual languages, but also by the diversity of linguistic features.

In terms of the number of speakers, the languages in WNG range from a hundred thousand to just a few speakers. Overall, the languages of WNG have small numbers of speakers. Some languages in the Central Highlands region have a sizable number of speakers, such as Western Dani (180,000 speakers), Ekari or Mee (100,000 speakers). Most of the other languages have under 10,000 speakers, as illustrated in Table 3.

Of the 276 languages in WNG, 254 have fewer than 10,000 speakers. This number covers about 92% of the total number of languages, including those with

2. Papuan Malay is a Malay-based lingua franca used throughout WNG. It is a pidginized language which has developed through a nativization process over about 200 years (Donohue 2011) and now has both native speakers and second language speakers throughout WNG. Kluge (2014: 5) mentions that Papuan Malay now has about 1.1 million to 1.2 million speakers. Among languages in WNG, the language is thus considered to have the largest number of speakers. Studies on Papuan Malay have been done by different people: Moeliono (1963), Silzer (1979), Samaun (1979), Donohue & Sawaki (2007), Scott et al. (2008), Sawaki (2005, 2007), Saragih (2012), Kluge (2014).

Table 3. Statistics of the number of speakers of the languages of WNG

No	Ranges of number of speakers	Number of languages	Names of languages (among others)
1	Over 100,000	2 languages	Western Dani & Ekari (Mee) (Papuan)
2	50,000–99,999	1 language	Mid Grand Valley Dani (Papuan)
3	10,000–49,999	19 languages	Biak, Ambai (Austronesian), Moni, Nalca, Meyah, Nguda, Sentani, Tehit, Ngalum, Yali (Papuan)
4	5,000–9,999	22 languages	Iha, Wandamen, Waropen (Austronesian), Ketengban, Asmat, Awyu, Citak, Kamoro, Mpur, Silimo, Wano, Wolani (Papuan)
5	1,000–4,999	78 languages	Ansus, Irarutu, Marau, Maya, Pom, Roon (Austronesian), Abun, Anasi, Mbaham, Bauzi, Iwur, Kokoda, Sawi, Sempan, Tabla, Una, Sobei (Papuan)
6	500–999	39 languages	Bonggo, Busami, Kawe, Kowiai, Kuri, Maden, Mor, Munggui, Onin, Ormu, Papuma (Austronesian), Betaf, Bonggo, Demisa, Duvle, Eritai, Iau, Kaburi, Kais, Korowai, Lepki, Mawes, Sikaritai, Skou, Taikat, Tanahmerah, Tunggare, Wonggom, Yahadian, Yei (Papuan)
7	100–499	86 languages	Arguni, As, Bedoanas, Erokwanas, Legenyem, Liki, Masimasi, Meoswar, Tobati, Urungnirim, Waigeo (Austronesian), Doutai, Elseng, Fayu, Kaiy, Keder, Kofei, Manem, Molaf, Yamna, Yelmek, Yoke (Papuan)
8	Below 100	29 languages	Anus, Duriankere, Dusner, Kayupulau, Mapia, Tandia (Austronesian), Awera, Benerif, Iresim, Itik, Kanum Badi, Kapori, Kembra, Kehu, Mander, Massep, Marori, Narau, Tangko, Tofomna Usku, Waria (Papuan)

fewer than 500 speakers. These languages are spread throughout WNG. Only 22 languages, or 8%, have more than 10,000 speakers. These languages are predominantly Papuan, with only two Austronesian languages, Biak and Ambai, in this category.

The geographical distribution of the languages is obvious. The languages with large numbers of speakers are mostly in the central highlands, such as Yali, Western Dani, Mee, Nduga, Central Great Valley Dani, and Mek, which have more than 10,000 speakers. In the southern region, languages such as Asmat, Marind, Kamoro, Awyu, Citak, and Muyu are considered languages with a medium number of speakers (fewer than 10,000). Except for Sentani, which has 30,000 speakers, the languages of the north coast, around the Jayapura and Lakes Plain areas,

are also small in terms of the number of speakers. Each of them has fewer than 1,000 speakers. In Cenderawasih Bay, which is home to most of the Austronesian languages of the West New Guinea group, most speaker populations are small, with the number of speakers ranging from 7,000 to less than 500. Some languages, such as Tandia (2 to 10 speakers) (Sawaki 2016b), Dusner (3 to 6 speakers) (Dalrymple & Mofu 2012), and Doreri (fewer than 10) (Sawaki 2022), a dialect of Biak, are endangered. As exceptions, two sizable languages are Biak (30,000 speakers) and Ambai (10,053 speakers), the two largest languages among the Austronesian languages. Speaker numbers for languages in the Bird's Head region and the Bomberai Peninsula are considered small. However, languages such as Hatam, Maybrat, Tehit, Meyah, and Sougb/Manikion are considered medium (more than 10,000 speakers).

This geographical distribution of languages, including the number of speakers, has changed over time, mainly due to the local and regional socio-political dynamics in WNG. Migration, both among native Papuans from various linguistic backgrounds and non-Papuan migrants from outside Papua, is growing. Many native Papuans are migrating from their traditional territories and resettling around big cities throughout WNG. Likewise, many non-Papuan migrants are also living in the big cities and other growing towns. This has resulted in demographic changes, as well as the socio-political roles of both native Papuans and immigrants from other parts of Indonesia in WNG, which have implications for sociolinguistic conditions. The life of indigenous Papuans in WNG can be sociolinguistically grouped into two categories based on settlement areas: Papuans living in big cities and those living in remote areas, primarily in villages. The demographic composition plays an important role in considering changes in languages and numbers of speakers in indigenous Papuan groups.

Sawaki and Arka (to appear) state that this sociolinguistic context is an example of how complex social, cultural, political, and economic pressures significantly affect the linguistic conditions in WNG. The dynamic development of WNG, starting from the era after World War II and culminating in the Special Autonomy era since 2000, has made cities such as Jayapura, Manokwari, Sorong, Merauke, Nabire, and Timika into destinations for non-Papuan migrants due to economic motivations. Note that these large cities are growing centres of government and economy. The population of these cities is now very diverse and multilingual, with indigenous Papuans from different linguistic backgrounds and non-Papuan migrants coexisting; some cities are even dominated by migrants (Pona 2008: 54), with Papuans becoming a minority. Economic, social, and political factors contribute to the marginalization of the native languages of Papua. In terms of language use, the presence of these migrants has meant that local languages are no longer spoken thoroughly, but are replaced by the increasing use of Indonesian

and local colloquial varieties of Papuan Malay, and, to a certain extent, English (Sawaki & Arka, to appear). The increase in the use of such languages to accommodate social, economic, and political needs is obvious. The Papuans, thus, are no longer able to practice the use of local languages because there are no Papuan settlements based on traditional cultural- and/or family-related systems (Budiono & Luthfiah 2019: 5; Mofu 2024; Sawaki 2022). They are away from their language communities for a long period of time and are creating a new generation, a city-born Papuan generation, without the ability to speak their parents' or grandparents' language. Mofu (2024) mentions that Manokwari-born children of the Pom-speaking community, whose parents/grandparents have resided and lived in Manokwari since the 1960s, do not speak Pom anymore. Likewise, Sawaki (2022) shows that almost all children and the young generation of the Doreri-speaking community around Manokwari are unable to speak their parents' or grandparents' native language due to being exposed to modernization and a multilingual environment.

In the rural areas in WNG, many places still retain their positive linguistic ecologies for native languages. Many languages, albeit with small numbers of speakers, are in areas geographically isolated by rivers, swamps, dense forests, mountains, and small islands, and are culturally vibrant, making them still safe and strong. Cross-generational speakers still use the language fluently. Languages like Bagusa (300 speakers), Bauzi (1,500 speakers), Burmeso (250), Burumakok (40), and others are still vibrant because they are geographically isolated (see Gordon 2005 for speaker numbers). Thus, these languages are still used in all domains by speakers of all ages. Furthermore, languages with sizable numbers of speakers, such as Western Dani, Yali, Asmat, Muyu, and Damal, are still vibrant because they are culturally strong. The languages are recognized as the social and cultural identity of the people, and these areas have also not experienced much change because development and accessibility to these areas are still lacking.

There are also language communities that are geographically quite isolated but, due to development, have experienced progressive change, especially in the world of education, with the establishment of schools and government offices that have existed from the early days of the Indonesian government. Many young people have experienced degradation in the use of local languages, as happened with Yaben speakers in Konda village. Many young people who still have local language knowledge prefer to communicate in Papuan Malay and even Indonesian (Sawaki 2016a; Mayor 2019; Karesina 2019). This is also the case with the Kais language in Kais village and the Waili language in Kapatlap and Wailebet villages, Salawati Island (Isnaeni et al. 2025). Many children have lost their ability to speak in their native languages.

Today's sociolinguistic conditions provide opportunities for changes in the existence of local languages in WNG. Many languages have shifted from a large number of speakers to a small number due to changing patterns of language use. This is especially true for the younger generation in many local language communities in WNG. The development of modernization, the growth of new settlements in both urban and rural areas, and the multilingual composition of these settlements, overlook the importance of local languages, and these factors have a huge impact on the existence of local languages in WNG.

Local governments at the provincial and regency levels have made efforts to protect and promote local languages in public and formal environments, such as in schools and offices. The Papua special autonomy law, article 58, obliges the WNG government to protect, preserve, and develop indigenous languages and culture in WNG, as several governments at the regency level have also stated in their local governmental regulations. This seems to be optimistic for indigenous languages in WNG to have a positive environment for language maintenance. However, the political will of the governments is still weak in this regard, and the implementation of these regulations is still unclear. They have not yet had a significant impact because the efforts made have not been maximized to create an intact environment for the use of local languages. The government regulations for the use of local languages are not supported by a strong commitment to planning and practicing local languages in intensive, systematic, and sustainable ways in both informal and formal environments.

3. A brief history of linguistic research and language documentation in WNG

This section describes the history of linguistic research and classical language documentation,[3] dating back to the presence of Western influence in WNG, i.e., the early spread of Christianity and the Dutch government. Missionaries had to learn local languages, as they were an effective means of communication for spreading Christianity. They then produced products such as word lists, sentence

3. In contrast to modern language documentation, following the idea of Himmelmann (1998, 2006), Arka (to appear) argues that classical or traditional documentation, in old days, refers to the type of documentation typically done by means of pen and pencil, without rich multimedia datasets and without digital technology. Although there were some audio and video recordings, they did not follow the procedures of modern language documentation, i.e. no data processing. The main deficiency of classical documentation is that its non-digital outcomes (e.g., text corpora) were not processed using the current technology of modern documentation.

lists, and short story forms and used them for Christian propagation efforts. Kamma (1981: 119) describes the first effort made by Carl Ottow, the first missionary in WNG, and J. G. Geissler, who made a word list of the Numfor language, a dialect of Biak, in 1857. This 1,000-word list of the Numfor dialect is probably one of the first products of local language documentation in Papua ever written. From 1855 to the 2000s, many works of language description and language documentation were produced by missionaries from various Protestant and Catholic churches throughout WNG. Most of these are written works in the form of word lists, sentence lists, written grammars, stories, songs, local language bibles, and local language literacy teaching materials.

The most intensive and focused research on language studies is that of the Summer Institute of Linguistics (SIL), a non-profit organization based in Dallas, Texas, USA, which worked in Papua from 1975 until about 2012. Over its 37 years of working in WNG, a great deal of linguistic research and language documentation was done. In each of their working areas, SIL workers conducted sociolinguistic surveys, grammar-writing, bible translation, and the development of literacy materials for teaching purposes. Note that SIL is no longer working officially in Papua today, and access to SIL materials is very limited and difficult.

During the Dutch government era in WNG, from 1949 to 1962, Dutch government officials were also assigned to study and document the local languages and cultures. Works by Dutch government officials included those of J. G. Held on the Waropen language (1942), K. W. Galis (1955) on languages and dialects in Netherlands New Guinea, H. J. K. Cowan (1965) on the Sentani language, J. C. Anceaux (1965) on the Nimboran language, and others. Note that these works were written in Dutch and only a few of them have been translated into English or other languages (Timmer & Ballard 2003). In the 1990s, the ISIR (Irian Jaya Studies Programme), an interdisciplinary research programme of several Dutch universities in collaboration with the Indonesian Institute of Sciences (LIPI) and Cenderawasih University, Jayapura, conducted intensive linguistic, cultural, and social research in the Bird's Head region of Papua; the results of this research can be accessed today (de Vries 2004; Mansoben 1995; Miedema 1997; Reesink 1995).

Since the region has been under Indonesian rule (i.e., since 1962), research in the field of language initiated by the Indonesian-based government and universities has declined dramatically. Especially in the early years, neither the government nor universities in Indonesia and Papua paid much attention to conducting research and documenting language and culture. Despite the interest of some local researchers at the University of Cenderawasih and later at the Universitas Papua, as well as the Jayapura Language Centre office, the results are still few. The Indonesian government focused more on politics, economics, and development. The main investments were in infrastructure, economy, and development. Cul-

tural and language development has not been a priority, so there is no clear vision, strategic plan, or long-term programme for language and culture preservation and development.

However, since 2000, several foreign researchers and a handful of Indonesian researchers, particularly in Papua, have conducted intensive research in WNG in collaboration with SIL, LIPI, which later became Badan Riset dan Inovasi Nasional (BRIN), Cenderawasih University, and the Universitas Papua. In particular, when the Center for Endangered Languages Documentation (CELD) at the Universitas Papua, which will be described more fully in Section 4, was established in 2009, some detailed language studies and modern language documentation work began, such as research on Biak[4] (van den Heuvel 2006; Mofu 2008), Wooi (Sawaki 2016a), Wamesa-Wandamen (Gasser 2014), Ambel (Arnold 2018), Mbaham (Cottet 2015), Roon (Gil 2017), and Yali (Riesberg 2021). Other researchers are conducting research independently, e.g., on Wano (Burung 2018), Kalamang (Visser 2022), Marori (Arka 2016), and others. Among these languages, Himmelmann has focused on working intensively on Wooi, Iha, Mpur, and Papuan Malay as part of language documentation activities at CELD. Several papers have been published to describe the different linguistic features of these languages, and a large amount of audio-visual documentary data has been archived in language documentation archives worldwide.

Interestingly, there have been hundreds of theses by students in the English department, Faculty of Literature and Culture, Universitas Papua, on the languages of Papua. Since 2006, students have been conducting language research in WNG. A wide range of topics, from phonology, morphology, syntax, semantics/pragmatics to discourse analysis, sociolinguistics, and ethnolinguistics, are within the scope of the students' theses. Note that the Universitas Papua is the only university in Papua that requires students of English to write their B.A. theses on the languages of Papua.

The historical works of linguistic description and classical language documentation became essential to language and culture documentation from the early period of Western influence in WNG. This period was dominated by pen and pencil documentation, although some recordings were produced. Yet, many products remain valuable today as references capturing a history of linguistic and classical language documentation in WNG.

---

4. The documentation of Biak (Heuvel 2006; Mofu 2008) was done before CELD was established. Heuvel (2006) compiled a corpus of Biak spontaneous speech in conjunction to his PhD thesis on Biak grammar. Likewise, Mofu (2008), together with his PhD supervisor Mary Dalrymple of the University of Oxford, documented Biak and produced a large corpus of Biak which is archived at the University of Oxford and in PARADISEC.

4. Modern language documentation in West New Guinea — a new perspective

Modern language documentation is also known as language documentation or documentary linguistics[5] (Himmelmann 1998, 2004, 2006; Lehmann 2001; Austin 2010; Grenoble 2010; Woodbury 2003, 2011). This field was introduced to Indonesian linguists in Indonesia, including WNG, in 2006[6] (Florey & Himmelmann 2007; Mayani & Katubi this volume). As a result, some language documentation projects have been conducted throughout Indonesia, including WNG. However, modern language documentation in Indonesia is mostly done by foreign researchers due to complex issues (Sawaki & Arka 2018). As a new field in language studies in Indonesia, most language institutions, including universities, national language agencies, national research and innovation agencies (BRIN), linguists, and language practitioners, lack theoretical, methodological, and practical knowledge about this field. Consequently, there are only a few Indonesian linguists, including the authors, who are well-trained in modern language documentation and involved in this field.

Overall, modern language documentation in Indonesia has been carried out in different regions by different people and funding agencies. Here, we list larger, multi-member projects that have been carried out in WNG, funded by one of the major language documentation funding agencies.[7] Modern language documentation in WNG began in 2009 with Nikolaus Himmelmann's and Yusuf Sawaki's Wooi documentation project (2009–2014), funded by the Volkswagen Foundation's *Dokumentation Bedrohter Sprachen* (DoBeS) programme, and conducted in close cooperation with CELD. In conjunction with the Wooi project, there

5. This is a subfield of linguistics that has its own methods, tools, and theoretical foundations for collecting and archiving lasting multipurpose records of a natural language (Himmelmann 2006; Austin 2014). It is also, by its nature, multi-disciplinary, which is not restricted to methods from linguistics but also from other fields such as ethnography, psychology, computer science, recording arts, and others (Himmelmann 1998; Harrison 2005; Coelho 2005).

6. Modern language documentation was introduced to Indonesian scholars when Nikolaus Himmelmann and some foreign scholars ran two workshops on language documentation in Ubud, Bali in 2006 and 2007. These workshops were considered as a milestone in introducing modern language documentation to the Indonesian scholars who were participants at the workshops.

7. Note that, of course, much more valuable documentary work has been done in the region, including in collaboration with CELD. In particular, there have been a number of PhD projects that have produced reference grammars and modern language documentation corpora. Many of these corpora are archived and publicly available in, e.g., ELAR or PARADISEC. Examples are Laura Arnold's work on Ambel, Emily Gasser's work on Wamesa, and Mae Carroll's work on Yei, among many others.

was the satellite project on Iha (by Sutriani Narfafan and Emmanuel Tuturop),[8] supervised by Nikolaus Himmelmann. The Wooi project was followed by another DoBeS-funded project, the 'Summits' project (2012–2016, under the lead of Sonja Riesberg and, again, Nikolaus Himmelmann), which was concerned with the Central Mountains languages Yali, Eipo, and Yale. The 'Summits' project also gave out scholarships to local students to conduct smaller-scale documentation projects. These led to the documentation of Western Dani (by Simon Tabuni),[9] Walak (by Peni Togodli), Mpur (by Boas Wabia),[10] and Wamesa-Wandamen (by Nova Refwalu). The joint documentary work between CELD and the team from the University of Cologne on Wooi, Iha, Mpur, and Yali is still ongoing until today.

Furthermore, a grant to document the Yaben and Kais languages was awarded to CELD by the Endangered Languages Documentation Programme (ELDP) (funding period 2019–2021). All these projects produced rich multimedia corpora (described in detail in Section 5.2.1), with publicly available online archives in the DOBES Archive and the Endangered Language Archive (ELAR). The data is also stored in the CELD Archive.

Similar language documentation projects have also been done in WNG, such as the ELDP-funded project, Marori and Smerky Language Documentation, which was carried out by a team from ANU led by I Wayan Arka, working with a trained member of the Marori-speaking community in Merauke. This project aimed to document the Marori and Smerky languages, which are categorized as extremely endangered. In the southwest region, there is a language documentation project on the Mor language conducted by Harald Hammarström.

Of all the language documentation programmes in WNG, the language documentation conducted at Universitas Papua, Manokwari, WNG, is well institutionalized and follows common standard practices in the world of language documentation. This language documentation activity at the Universitas Papua also now has its archive system managed by the Center for Endangered Languages Documentation (CELD) at Universitas Papua. All corpora are also stored in several archive systems abroad. The archive system at CELD will be discussed in detail in Section 5.

The language documentation implemented in WNG not only includes the activities of recording and storing data, but also involves other activities such as training, workshops, and office management, including library management, computer and software skills specifically related to language documentation. These

8. Sutriani Narfafan no longer works with CELD, but Emmanuel Tuturop still does. He is still working on Iha.

9. Simon Tabuni used to work at CELD and is now doing a PhD at Uppsala University, Sweden.

10. Since his student project in 2014, Boas Wabia has worked as a member of staff at CELD.

programmes are necessary because they are closely related to efforts to institutionalize language documentation. What has been done in WNG through CELD so far is an effort to build a long-term sustainable language documentation infrastructure in WNG.

5. Center for Endangered Languages Documentation (CELD): A strategy to establish a sustainable language documentation infrastructure in Western New Guinea

This section contains some discussion about the Center for Endangered Languages Documentation (CELD), the only language documentation institute or similar institution in WNG and possibly Indonesia. It is housed at the Universitas Papua in Manokwari, West Papua, Indonesia. This section discusses the history, vision, mission, and goals of CELD, the programmes and activities undertaken at CELD, and the future challenges of CELD and its activities.

5.1 CELD: Its history, vision & mission, and purposes

One significant challenge in WNG is to implement language documentation practices into the establishment of a sustainable language documentation programme. This challenge involves the development of a language documentation infrastructure to ensure that language documentation activities can be sustained over a long period. These efforts have been made in other places in Indonesia, but have been difficult to realize. After attending two workshops on language documentation in Bali in 2006 and 2007 (Florey & Himmelmann 2007; Mayani & Katubi, this volume), efforts to establish a language documentation institute began in WNG.

The first author of this article was a participant at the 'Bali workshops', where he began intensive discussions and came up with strategic steps together with his mentor Nikolaus Himmelmann and Alexander Loch, a development consultant, to write a joint proposal for funding from the Volkswagen Foundation in 2008, successfully securing funding for the Wooi language documentation programme (see Section 4). Through this funding, a language documentation institute, the Center for Endangered Languages Documentation (CELD), was established.

CELD was officially inaugurated at Universitas Negeri Papua[11] on February 9, 2009. The purpose of establishing CELD was very simple, namely, to document

11. Universitas Negeri Papua is a former name of Universitas Papua. It was still under Universitas Negeri Papua when CELD was first started. The name was changed from Universitas Negeri Papua to Universitas Papua in 2013.

the languages and cultures in Papua before they are lost. The argumentation was that language and culture are the identity of a language community, and if their language and culture are lost, their identity, self-esteem, and pride will also be lost. This is a reality in WNG where hundreds of languages exist with their rich cultural properties. We realized that most of the hundreds of languages in Papua will experience a decline in the number of speakers over time. Therefore, CELD was established with the following vision:

a. to work closely with language communities in WNG to document their languages and cultures;
b. to train and prepare local linguists, students, and local communities to use state-of-the-art language documentation techniques;
c. to support teachers, government agents, artists, and cultural and language activists in developing and using materials in local languages; and
d. to commit to establishing and developing a sustainable structure for accessing linguistic and anthropological data from around WNG at Universitas Papua.

The centre also has the following missions:

a. to be the centre for language documentation and conservation in WNG;
b. to be a long-term, sustainable centre in the field of language documentation;
c. to become a regional digital archiving centre in Indonesia and the Pacific;
d. to become a centre for multi-level capacity building for linguists, language communities, and university members, including lecturers and students; and
e. to become a centre for national and international collaborations.

CELD is led by the first author as director and supported by five staff members, all of whom are local staff with varied backgrounds. Some of them graduated from Universitas Papua, and some are members of local speech communities. They were recruited and trained not only to do language documentation activities in the field, but also in data management and archiving, computer skills, office management, library management, and accounting. All of this was done with a 'learning by doing' approach, where they were trained directly by running the Wooi language documentation programme. These staff members thus became proficient and qualified through their work at CELD.

The collaborative work and intensive supervision from Nikolaus Himmelmann and his team from the University of Cologne, Germany for the CELD staff effectively developed the capacity of the centre, including human resources and management. Ongoing supervision is carried out both in person at the CELD office in Manokwari and online from Germany. The process of mentoring and supervision is carried out in both fieldwork and office work. This 'learning by doing' approach, through continuous supervision and mentoring, has increased

the theoretical and practical knowledge of language documentation and other office work of CELD staff. There is testimony that a young and talented man from Wooi village, named Jimi Kirihio, who had only finished high school and had lived in the village all his life, was recruited to work at CELD. He was trained and mentored to do documentation work in his language, the Wooi language. He was a very fluent Wooi speaker, but he was unfamiliar with computers, data management, office work, and other related aspects. Within a few years, however, he became very advanced and today has a lot of knowledge in the field of language documentation, data management, general linguistics, and office management. He is also providing training to students at the Universitas Papua on language documentation techniques.

This is what it takes to build a sustainable infrastructure in language documentation: to increase the capacity of human resources and develop the necessary knowledge in the world of language documentation. It is believed that human resources are the crucial factor for sustaining the vision at CELD. Therefore, all subsequent CELD staff recruited also went through the same capacity-building process. Once these CELD staff become capable, they are independently able to operate and manage the institution and its activities. They are also ready to train and mentor students and others learning about language documentation.

5.2 CELD programmes and activities

5.2.1 *Language documentation projects*

CELD has worked on several language documentation projects in Papua. The projects are the above-mentioned DoBeS projects Documenting Wooi (2009–2014) and Documentation Summits in the Central Mountains of Papua (2012–2016), and the ELDP-funded Konda and Kais Language Documentation Project in the southern Bird's Head region (2019–2021). There was also a project documenting the lexicons of the languages of Papua, funded by the Max Planck Institute (MPI) in Jena, Germany (2018–2021). These projects are called major projects because, through these projects, other satellite documentation activities are also carried out for other languages, as well as other programmes in CELD. Through these projects, CELD has been able to store and archive large corpora of the documented languages; see Table 4.[12]

Within the MPI-funded project, CELD collected and recorded 251 culture-based words in individual languages. Thus far, there are 26 languages in WNG

12. Note that this table only contains data collected in projects under CELD lead or with the active participation of CELD staff. It does not include the data collected and deposited by visiting researchers and PhD students sponsored by and collaborating with CELD.

Table 4. The list of languages and the size of corpora documented and archived at CELD

No	Language	Total files (approx.)	Types of corpora	Total bytes	Total hours (approx.)
1	Wooi	500	Audio, video	172.9 GB	70
2	Iha	300	Audio, video	138.9 GB	50
3	Eipo	90	Audio, video	39.2 GB	1
4	Yali	690	Audio, video	236 GB	95
5	Walak	260	Audio, video	61.4 GB	40
6	Western Dani	30	Audio, video	15.32 GB	5
7	Wamesa	40	Audio, video	4.25 GB	5
8	Yale Kosarek	70	Audio, video	10.1 GB	1
9	Mpur	160	Audio, video	9.71 GB	20
10	Yaben (Konda)	320	Audio, video	439 GB	35
11	Kais	120	Audio, video	1370 GB	120
	Total	**2580**		**2,376.78 GB**	**442**

that have been recorded. There are about 11 hours of recordings in total. The corpora are archived both at CELD and at MPI Jena.

5.2.2 *Trainings, workshops, and academic courses*

CELD also functions as an institution that conducts training and workshops on language documentation and linguistic description (Gasser 2013, Riesberg to appear, Sawaki et. al. 2016). Training and workshops are a core mission of CELD (see above), which aims to increase and strengthen human capacity in the field of language documentation. The targeted participants are lecturers, students at the Universitas Papua campus, and language community members in WNG. Some of the workshops and training activities carried out by CELD include:

1. Training and workshops for students and lecturers in 2010, 2011, and 2012, conducted within the framework of the "Documenting Wooi" project.
2. Training and workshop for members of language communities in 2013 within the framework of the project "Documentation Summits in the Central Mountains of Papua" for the Yali, Eipo, and Yale Kosarek languages.
3. Training and workshops for students and lecturers in 2014 and 2015, conducted within the framework of the project "Documentation Summits in the Central Mountains of Papua." Note that this training contributed to the success of the associated projects listed in Table 4.

4. Training and workshop for students and lecturers in 2023, conducted within the framework of the CELD grant for capacity building.
5. Training for lecturers and students on language documentation and dictionary making in 2025.

In addition to training and workshops, CELD also aims to support the language documentation and conservation course within the curriculum of the English Literature study programme, Faculty of Letters and Culture, Universitas Papua. Since 2008, the language documentation and conservation (LDC) course has been integrated into the curriculum; this course has 2 credits and covers the purposes of the language documentation, the background to language ecology in WNG, and the theory, methodology, and practices of language documentation. The course aims to provide students with theoretical and practical knowledge of language documentation.

One of the strengths of this course is the direct practice of conducting language documentation with native speakers of languages in Papua. Luckily, Universitas Papua has students from many different language backgrounds, and they can contribute to this course as language consultants. Thus, each year, we can easily manage projects on language documentation for the course, such as the documentation of the Yali language and the documentation of the Arandai language, Tomu dialect.[13] By including language documentation as a course in the Faculty of Letters and Culture, the language documentation programme is institutionalized and will be long-lasting. For the past 16 years, this course has been conducted by CELD, with its staff providing the teaching at the faculty.

5.2.3 *Library services*

Another service provided by CELD is a growing library in linguistics, language documentation, and anthropology. This comprehensive language and culture specialty library at Universitas Papua is perhaps the most comprehensive one of any university library in WNG. The purpose of establishing this library is to provide references in the fields of language, language documentation, and culture, not only to students and lecturers at Universitas Papua, but also to the wider community.

The reading collection in the CELD library consists of two forms, namely, hard copies and digital copies. The hardcopy collection of books in this library has been donated by various parties from around the world. Linguists and anthropologists bring books directly when they visit the CELD or send them via the post office. In addition to the hard copies, there are also thousands of digital copies collected by CELD staff and donated by CELD partners.

13. See https://www.youtube.com/watch?v=TEcIqlvg3h8 and https://www.youtube.com/watch?v=uakkIRIjf1o for examples.

In addition to the book collection, the CELD library also has a collection of student theses, specifically linguistics theses. There are around 200 theses ranging from sociolinguistics, phonology, morphology, syntax and semantics/pragmatics to ethnolinguistics and other subjects. These theses are the original work of students describing the languages of Papua. The library metadata is available to help users find the specific reference they need. Note that in the long run, these theses will be included in the open access system,[14] so that they can be read by many people. The library is visited by students almost every day, especially those who are writing their theses. It is also often used by researchers from Indonesia and abroad who visit CELD.

5.2.4 *Center for national and international collaboration*

It is part of its philosophy that CELD will grow professionally by building collaboration with various parties, both within Indonesia and beyond. There have been numerous collaborations, both to develop the capacity of CELD and to support language and culture research in WNG. Collaboration to increase the capacity of CELD has been carried out since CELD was established, and has become a force with long-term goals. A long-standing collaboration for capacity building purposes is with the University of Cologne, Germany. The university has a long-term commitment to support CELD with funding, projects, and technical advice. Nikolaus Himmelmann and Sonja Riesberg have visited CELD every year since its establishment (with a short break during the COVID pandemic). Several of their students and staff have also visited CELD for both completed projects and ongoing projects run between CELD and the University of Cologne.

Since its establishment, CELD has collaborated with several international partners such as the Max Planck Institute for Evolutionary Anthropology (MPI-EVA) in Leipzig, MPI Jena, the Australian National University (ANU), and, more recently, Uppsala University, Sweden. At the national level, CELD collaborates with the National Research and Innovation Agency, Jakarta, in documenting Waili (2022) and Mairasi (2024), as well as with the Linguistic Society of Indonesia (MLI) and the National Language Agency.

In addition to the institutions mentioned above, cooperation is also carried out with several individual foreign researchers from the Netherlands, Australia, the United States, Japan, and some from universities in Indonesia. This cooperation includes several activities, namely:

---

14. CELD gets funding from the MPI, Leipzig to digitalize all theses at the English Department, Universitas Papua; they are stored in the archive system at CELD and at the language database at Tokyo University for Foreign Studies. This will be an open access system under certain circumstances.

1. Research
2. Training
3. Seminars and conferences

In terms of research collaboration, CELD stresses 'the shared-knowledge approach', in which foreign/national visitors are expected to work together with CELD staff, students, lecturers and members of the language community. This approach forms a shared knowledge triangle through which the knowledge is transferred.

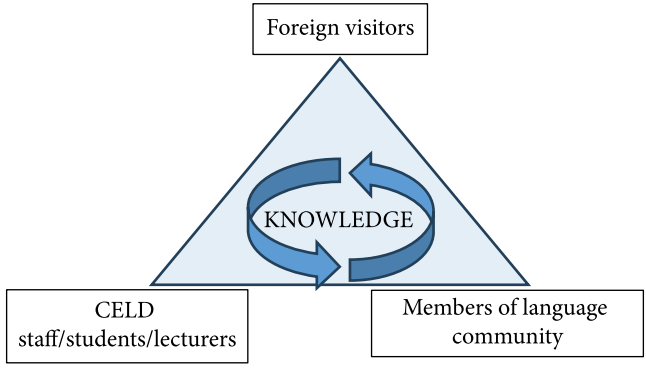


Figure 1. The "shared-knowledge approach" represents the collaborative activities at CELD

Through this approach, CELD believes that the three parties create a channel for knowledge and experience to be transferred from one party to another, and create an opportunity for each party to increase their capacity; the visitor gains knowledge about the language, culture, and social richness of Papua; the CELD staff, students, and lecturers gain knowledge about certain theories and practices applied in academic circumstances, while also gain knowledge about languages, cultures, and social life from the language community members. The language community members also gain knowledge from both visitors and CELD.

With strong technical and financial cooperation and networks, CELD has also become the venue for international seminars and conferences in Manokwari. One international conference sponsored by CELD and Universitas Papua is the Workshop of the Languages of Papua (WLP), which has been held six times in Manokwari. This workshop is the only international conference held in WNG so far that focuses on Austronesian and Papuan languages of New Guinea and the adjacent islands. Many linguists have visited CELD in Manokwari during this conference. In addition to the WLP, the Universitas Papua and CELD, as co-organizers, also hosted the 2018 International Conference of the Indonesian Lin-

guistic Society (KIMLI) in Manokwari. Many participants from other parts of Indonesia and from several other countries attended this conference in Manokwari. The purpose of these conferences is not only to share products of linguistic and cultural research, but also to build collaborations and networks with other researchers.

6. Several challenges faced by CELD

The fundamental questions when looking at the challenges for CELD in the future are: How can CELD survive in the future? Will CELD still exist ten years from now? These questions describe the challenges faced during CELD's 16-year journey in WNG, and are concerned with whether CELD can continue to carry out language documentation activities in the future. Beyond the good and successful stories about CELD, the organization has also faced challenges in carrying out its language documentation mission. There are four challenges that CELD faces in the future for running language documentation programmes in WNG, namely:

a. human resources,
b. long-term & short-term funding,
c. training efforts, and
d. national/international collaboration.

Note that the four challenges mentioned are not mutually exclusive but partly intertwine with each other.

6.1 Human resources

In Indonesia, becoming a government worker is still a high preference for most people. Jobs such as becoming a language documentarist have a low priority. People seek jobs that provide a future guarantee in their lives, earning a better salary. Thus, only a few people are interested in working at CELD. This gives CELD a problem with human resources in the future. Luckily, CELD still has a strong team, consisting of five permanent staff members, who manage and organize the work at CELD.

Since its establishment, approximately 15 people have worked at CELD. Of these 15 staff members, only five are currently still working at CELD. Most other staff have had to leave because their work was project-based. After their projects were completed, they left CELD and found work elsewhere. This resulted in a reduction of staff at CELD. For this reason, CELD has taken steps to retain a

small number of permanent staff and will contract other people when projects are implemented. In addition to CELD staff, several international partners regularly visit and support CELD activities by providing advice and technical guidance. To date, the international partners collaborating with CELD and financially supporting CELD staff are Nikolaus Himmelmann and Sonja Riesberg.

Human resources development at CELD is very much dependent on the projects and funds available. CELD is always developing a strategic plan to start projects that will provide opportunities for students and community members to be employed by CELD. CELD has not yet become a prosperous place to get a permanent job. Therefore, only with competitive and attractive projects will it be easy for CELD to find people to work here. In addition, while projects always provide funding that can be used for the staff's salaries, the CELD staff salaries depend entirely on the project's funds. With competitive and reasonable salaries, it will be possible for CELD to offer a long-term perspective to its well-trained staff members.

6.2 Long-term and short-term funding

Funding is a crucial element in the operation of CELD. Language documentation activities in Indonesia, in general, and in WNG in particular, are activities that until now have been difficult to get funding for from the government and universities, although national and local regulations state that language and culture preservation is an obligatory mandate of the state government. In WNG, the Special Autonomy Law obliges the provincial governments in Papua to protect and develop Papuan cultures and languages. Consequently, the government must provide funding for activities relating to protecting and developing Papuan cultures and languages. However, in practice, this is not a priority in developmental programmes. CELD, in contrast, has been run for 16 years with foreign funding. Language documentation projects provide most of the funding for CELD activities, such as training, salaries, travel, fieldwork, and office activities. In addition, another strategy to get funding is to require foreign visitors whose visas are sponsored by CELD to pay a sponsorship fee. This, however, cannot cover most expenses at CELD. Some field research activities independently pursued by CELD have also brought in funds that have been used for the benefit of CELD, but these activities have been short-term and sporadic.

Other attempts to secure funding from the government and universities have been made, but have not been successful in securing long-term funding that would ensure the continuation of CELD activities. CELD therefore still needs funding support from abroad. Collaboration with international partners is still needed to carry out all activities in CELD.

6.3 Training efforts

To increase participation from students, lecturers, and members of the language communities in WNG in language documentation activities, there is a need for strategic and well-developed training. The purpose of the training is to produce well-trained actors who can participate in language documentation activities in WNG. CELD has done several training sessions and workshops thus far. With the skills acquired through well-designed and continuous training and workshops, participants can also conduct their own language documentation activities (usually for members of the language communities) and try to apply for research and language documentation grants offered by grant providers.

So far, most participants in our training and workshops have been students and lecturers at the Universitas Papua There are not many participants from among language community members. This is due to limited funds, so CELD cannot reach out to community members to attend trainings and workshops. For this reason, in the future, CELD will need to seek funds earmarked directly for training and workshops for community members. This is also true for training students and lecturers, who also require funding so that CELD can run training continuously.

6.4 National and international collaboration

As mentioned above, collaboration is a significant factor in improving the work of CELD, both in terms of quality and quantity. Collaborations have contributed to the development of human resources, funding, management, and the organization of CELD, as well as the corpora archived at CELD. Important issues in maintaining collaborations between CELD and other institutions, i.e., universities, funding bodies, and individual scholars, are the following:

a. Strong collaborations form social networks between different parties through which resources can be channelled to CELD, allowing language documentation work to become more effective, efficient, professional, and of a high standard.
b. Collaborations also contribute to professional development, where CELD staff are exposed to new tasks that require new skills. The projects offered often use new methods, skills, and technologies. CELD needs to do specific training to keep up with changing expectations. Thus, CELD staff must increase professional development to fulfil the goal.
c. Collaborations also create a lively and dynamic scientific environment with a mutual exchange of knowledge among CELD staff and foreign/national scholars. This is a challenge for the future to maintain performance with high

standards in language documentation, especially with developments in the world of language documentation in terms of theory, methods, and technology. The same standard procedure is also applied for the archive system in which the corpora stored in CELD must be of the same standard as in other archive systems.

d. Collaborations provide an opportunity for CELD to obtain funding that can help it to carry out language documentation work in WNG, which has many languages identified as endangered. Considering the actual conditions in Indonesia, particularly in WNG, where government and other funding commitments pose challenges to conducting language documentation, collaborations with international partners will always be a valuable contribution to CELD's continued work in the language documentation field.

7. Final remarks

In this final section, we reflect on our experience running CELD since 2009, and on its vision, missions, and goals. CELD is not just a centre for academic purposes. It has become the centre of a scholarly environment, where knowledge, skills, and experience are shared between foreign scholars, local scholars, and members of the language communities. It is also a place where other issues, such as academic habits and social and cultural differences, are neutralized for the success of the language documentation activities. In some ways, there are cultural and social gaps between foreign scholars and local people in WNG. We have been aware of the different expectations brought by foreign linguists, local linguists, and language communities while working in WNG. Foreign linguists work with advanced knowledge and high-standard working habits and procedures. Foreign linguists carry this academic habit, which has become a solid foundation to strengthen the capacity of CELD. They can also compete to obtain grants from different funding agencies, such as the Endangered Language Documentation Programme (ELDP), National Endowment for the Humanities (NEH), The Endangered Language Fund (ELF), etc., because of their experience and expertise in applying for the grants.

Local linguists may not have a solid academic background in linguistics and modern language documentation. Thus, they are not well-prepared to carry out fieldwork and to work with theory and technology in language documentation. However, they have a solid foundation in the linguistic, social, political, and cultural backgrounds of WNG, which can contribute to strong and successful communication with language communities. Some of them even have rich fieldwork experience in studying languages and cultures in the region. They may become

strong partners for the foreign linguists to do fieldwork with in WNG, where they gain knowledge and skills from the training and workshops facilitated by the foreign linguists. Since they are native Papuans, they become actors in building relationships and networks among language communities by using their local knowledge of the people. This is the best practice implemented by CELD.

Members of language communities are obligatory partners, being sources for their native languages and cultures. All insight and knowledge of a language and culture derive from the native speakers, and they are indispensable teachers of Papuan life, languages, and cultures. For the sake of language documentation, our experience in CELD confirms that involving members of the language community in the programme leads to successful outcomes. Most of our projects have involved bringing members of the language communities to work with us both in the field and in the office. They work closely with CELD staff in documenting their languages, such as Wooi, Iha, Mpur, and Yali. The members of the local community have been trained and have carried out fieldwork in and on their native languages. Their deep understanding of their languages and cultures becomes an essential key to the success of the CELD projects in WNG. Therefore, in many ways, we prefer to recruit members of local communities to work at CELD, rather than people from outside Papua.

To conclude, modern language documentation in WNG is still relatively new to both the academic institutions and the local communities. It therefore needs strong efforts to face both academic and non-academic challenges. These efforts, such as human resource development, long-term and short-term funding, training and workshops, and national and international collaboration, contribute to the sustainability of language documentation in WNG. CELD is a successful example of how an effort is locally institutionalized in WNG to support language documentation activities in the region, where nearly 300 languages exist.

References

Anceaux, J.C. 1965. *The Nimboran Language*. VKI 44. 's-Gravenhage: Martinus Nijhoff.

doi Arka, I Wayan. 2016. "Externally and Internally Headed Relative Clauses in Marori." In *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar, Polish Academy of Sciences, Warsaw, Poland*, ed. by Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King, and Stefan Müller, 23–42.

doi Arka, I Wayan. To Appear. "Multidimensionality of Capacity Building in an Indonesian Ethno-Ecological Context." In *Projecting Voices: Essays in Language and Linguistics in Honour of Jane Simpson*, ed. by Carmel O'Shannessy, James Grey, and Denise Angelo. Canberra: Asia-Pacific Linguistics.

Arnold, Laura. 2018. A Grammar of Ambel: An Austronesian Language of Raja Ampat, West New Guinea. Doctoral Dissertation, University of Edinburgh. https://era.ed.ac.uk/handle/1842/31120

Austin, Peter K. 2010. "Current Issues in Language Documentation." *Language Documentation and Description* 7: 12–33. London: SOAS.

Austin, Peter K. 2014. "Language Documentation in the 21st Century." *JournaLIPP* 3: 57–71.

Barclay, Peter G. 2008. *A Grammar of Western Dani*. Munich: Lincom.

Bellwood, Peter. 1997. *Prehistory of the Indo-Malaysian Archipelago*. Honolulu: University of Hawai'i Press.

Berry, Keith. 1995. A Description of the Abun Language. Phonology and Basic Morphosyntax. MA Thesis, La Trobe University.

Blust, Robert. 2024. "Linguistic Approaches to Austronesian Culture History." In *The Oxford Guide to the Malayo-Polynesian Languages of Southeast Asia*, ed. by Alexander Adelaar, and Antoinette Schapper, 35–48. Oxford: Oxford University Press.

Bromley, Myron H. 1981. *A Grammar of Lower Grand Valley Dani*. Canberra: Pacific Linguistics.

Budiono, Satwiko, and Devi Luthfiah. 2019. *Revitalisasi Bahasa Tobati Berbasis Sekolah di Kota Jayapura*. Jakarta: Badan Pengembangan Bahasa dan Perbukuan.

Burung, Willem. 2017. A Grammar of Wano. Doctoral Dissertation, University of Oxford.

Burung, Willem. 2018. "Alienable and Inalienable Nouns in Wano." *Linguistik Indonesia*, 36 (1): 37–65.

Coelho, Gail. 2005. "Language Documentation and Ecology: Areas of Interaction." *Language Documentation and Description* 3, 63–74. London: SOAS.

Cottet, Fanny. 2015. A Phonology of Mbahám: Reduction and Contrast. Doctoral Dissertation, The Australian National University.

Cowan, H. K. J. 1965. *Grammar of the Sentani Language*. VKI 47. 's-Gravenhage: Martinus Nijhoff.

Dalrymple, Mary, and Suriel Mofu. 2012. *Dusner*. Munich: Lincom.

de Vries, Lourens. 2004. *A Short Grammar of Inanwatan, an Endangered Language of the Bird's Head of Papua, Indonesia*. Canberra: Pacific Linguistics.

Doble, Marion. 1987. "A Description of Some Features of Ekari Language Structure." *Oceanic Linguistics* 26 (1): 55–113.

Dol, Philomena. 2007. *A Grammar of Maybrat: A Language of the Bird's Head Peninsula, Papua Province, Indonesia*. Canberra: Pacific Linguistics.

Donohue, Mark. 2005. "Tone and the Trans New Guinea Languages." In *Proceedings of the Symposium 'Cross-Linguistic Studies of Tonal Phenomena: Historical Development, Tone-Syntax Interface, and Descriptive Studies'*, ed. by Shigeki Kaji, 33–54. Tokyo: Tokyo University of Foreign Studies. http://hdl.handle.net/1885/36384

Donohue, Mark. 2011. "Papuan Malay of New Guinea. Melanesian Influence on Verb and Clause Structure." In *Creoles, Their Substrates, and Language Typology*, ed. by Claire Lefebvre, 413–435. Amsterdam: John Benjamins.

Donohue, Mark, and Yusuf Sawaki. 2007. "Papuan Malay Pronominals: Forms and Functions." *Oceanic Linguistics* 46 (1): 253–276.

Drabbe, Peter. 1959. *Grammar of the Asmat Language*. Syracuse, Indiana: Our Lady of the Lake Press.

Evans, Nicholas, I Wayan Arka, Matthew Carroll, Yun Jung Choi, Christian Döhler, Volker Gast, Eri Kashima, Emil Mittag, Bruno Olsson, Kyla Quinn, Dineke Schokkin, Philip Tama, Charlotte van Tongeren, and Jeff Siegel. 2018. "The Languages of Southern New Guinea." In *The Languages and Linguistics of the New Guinea Area: A Comprehensive Guide*, ed. by Bill Palmer. 641–774. Berlin: De Gruyter Mouton.

Flassy, Don A.L. 1991. Grammar Sketch of Tehit, a Toror Language, the West Doberai Peninsula, New Guinea (Irian Jaya). MA Thesis, Rijksuniversiteit te Leiden.

Florey, Margaret, and Nikolaus P. Himmelmann. 2007. "New Directions in Field Linguistics: Training Strategies for Language Documentation in Indonesia." In *Endangered Languages of Austronesia*, ed. by Margaret Florey, 212–140. Oxford: Oxford University Press.

Foley, William A. 1986. *Papuan Languages of New Guinea*. London: Cambridge University Press.

Foley, William A. 2000. "The Languages of New Guinea." *Annual Review of Anthropology* 29: 357–404. https://www.jstor.org/stable/223425.

Galis, Klaas W. 1955. "Talen en Dialecten van Netherlands New-Guinea." *Tijdschrif Nieuw-Guinea* 16: 109–117, 134–145, 161–177.

Gasser, Emily. 2013. *Exchanging Words and Skills: Language Documentation in West Papua*. Presentation, 3rd International Conference on Language Documentation and Conservation, University of Hawai'i at Mānoa, Honolulu, 26 February — 1 March, 2013. https://hdl.handle.net/10125/26591

Gasser, Emily. 2014. Windesi Wamesa Morphophonology. Doctoral Dissertation, Yale University, New Haven.

Gasser, Emily, Laura Arnold, and David Kamholz. 2024. The Languages of Halmahera and West New Guinea. In *The Oxford Guide to the Malayo-Polynesian Languages of Southeast Asia*, ed. by Alexander Adelaar, and Antoinette Schapper, 629–651. Oxford: Oxford University Press.

Gil, David. 2017. "Roon *Ve*, DO/GIVE Coexpression, and Language Contact in Northwest New Guinea." In *Contact and Substrate in the Languages of Wallacea Part 1*, ed. by Antoinette Schapper, 41–100. Tokyo: ILCAA.

Gravelle, Gilles. 2010a. A Grammar of Moskona. An Eastern Bird's Head Language of West Papua, Indonesia. Doctoral Dissertation, Vrije Universiteit, Amsterdam. https://hdl.handle.net/1871/15730

Gravelle, Gilles. 2010b. *Meyah, a Language of West Papua, Indonesia*. Canberra: Pacific Linguistics.

Grenoble, Lenore A. 2010. "Language Documentation and Field Linguistics: The State of the Field." In *Language Documentation: Practice and Values*, ed. by Lenore A. Grenoble & N. Louanna Furbee, 289–309. Amsterdam: John Benjamins.

Harrison, K. David. 2005. "Ethnographically Informed Language Documentation." *Documentation and Description* 3: 22–41. London: SOAS.

Hartzler, Dwight. 1976. "A Study of Sentani Verb Structure." In *Irian: Bulletin of Irian Jaya Development*, 5 (2): 18–39. Jayapura: Cenderawasih University.

Held, Gerrit J. 1942. *Grammatica van het Waropensch (Nederlandsch Noord Nieuw Guinea)*. Bandung: Nix.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2004. "Documentary and Descriptive Linguistics." In *Lectures on Endangered Languages, Volume V*, ed. by Osamu Sakiyama, and Fubito Endo, 37–83. Kyoto: Endangered Languages of the Pacific Rim.

Himmelmann, Nikolaus P. 2006. "Language Documentation: What Is It and What Is It Good For? In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin: Mouton de Gruyter.

Holton, Gary, and Marian Klamer. 2018. "The Papuan Languages of East Nusantara and the Bird's Head." In *The Languages and Linguistics of New Guinea: A Comprehensive Guide*, ed. by Bill Palmer. 569–629. Berlin: Mouton de Gruyter.

Isnaeni, Mukhammad, Katubi, Dendi Wijaya, Ernianti, Jusmianti Garing, Andi Faisal, and Yusuf Sawaki. 2025. "Vitality of the Waili Papua in Salawati Island: What Makes Waili Papua's Language Endangered?" In *Some Efforts in Language and Literature Preservation in Indonesia. From Sumatra to Papua*, ed. by Mukhammad Isnaeni, Katubi, and Mulatamia R. M. T. Lauder. 237–264. Hauppauge: Nova.

Kamholz, David. 2024. "Historical Linguistics of the South Halmahera-West New Guinea Subgroup." In *The Oxford Guide to the Malayo-Polynesian Languages of Southeast Asia*, ed. by Alexander Adelaar, and Antoinette Schapper, 181–187. Oxford: Oxford University Press.

Kamma, Freerk C. 1981. *Ajaib di Mata Kita, Jilid 1*. Jakarta: BPK Gunung Mulia.

Karesina, Stevani F. C. 2019. A Study on Language Use in Yaben. B.A. Thesis, Universitas Papua, Manokwari.

Karubaba, Sara Yulita. 2017. "The Voice System of Papuan Malay: A Typological Study." *MELANESIA: Jurnal Ilmiah Kajian Sastra dan Bahasa*, 1 (2): 57–64.

Kluge, Angela. 2014. *A Grammar of Papuan Malay*. Utrecht: LOT.

Lehmann, Christian. 2001. "Language Documentation: A Program." In *Aspects of Typology and Universals*, ed. by Walter Bisang, 83–97. Berlin: Akademie Verlag.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig. 2014. *Ethnologue: Languages of the World*. 17th Edition. Dallas, Texas: SIL International.

Lilley, I. 1992. "Papua New Guinea's Human Past: The Evidence of Archaeology." In *Human Biology in New Guinea: The Small Cosmos*, ed. by Robert D. Attenborough, and Michael P. Alpers, 150–71. Oxford: Clarendon.

Mansoben, Johszua R. 1995. *Sistem Politik Tradisional di Irian Jaya*. Jakarta: LIPI-RUL.

Mayor, Infak Isaswar. 2019. Possessive Constructions in Yaben. B.A. Thesis, Universitas Papua.

McPherson, Laura, and Matthew Dryer. 2021. "The Tone System of Poko-Rawo (Skou)." *Phonological Data and Analysis*, 3 (1): 1–32.

Miedema, Jelle. 1997. *Texts From the Oral Tradition in the Eastern Bird's Head Peninsula of Irian Jaya*. Leiden/Jakarta: DSALCUL/ISIR.

Moeliono, Anton M. 1963. Ragam Bahasa di Irian Barat. In *Penduduk Irian Jaya*, ed. by Koentjoaraningrat, and Harsja W. Bachtiar, 28–38. Jakarta: Penerbitan Universitas.

Mofu, Stiffen Y. 2024. Language Use of Pom People Living in Amban Perkam. B.A. Thesis, Universitas Papua, Manokwari.

Mofu, Suriel S. 2008. Biak Morphosyntax. Doctoral Dissertation, University of Oxford.

Ndiken, Novita Y. 2011. Verbal Agreement in Malind. B.A. Thesis, Universitas Papua, Manokwari.

Ode, Cecilia. 1997. "A Descriptive Study of Prosodic Phenomena in Mpur (West Papuan Phylum)." In *Proceedings of the 5th European Conference on Speech Communication and Technology*, 183–186.

Olsson, Bruno. 2021. *A Grammar of Coastal Marind*. Berlin: De Gruyter Mouton.

Peckham, Lloyd. 1982. "Mairasi Verb Morphology." In *Workpapers in Indonesian Linguistics* 1, ed. by Marit Kana, 75–96. Jayapura: UNCEN-SIL.

Pona, La. 2008. "Penduduk, Otonomi Khusus, dan Fenomena Konflik di Tanah Papua." *Jurnal Kependudukan Indonesia* III (1): 51–67.

Reesink, Ger P. 1995. "Explorations in Bert's Head: Hatam." In *Tales From a Concave World: Liber Amicorum Bert Voorhoeve*, ed. by Connie Baak, Mary Bakker, and Dick van der Meij, 124–138. Leiden: Department of Languages and Cultures of South-East Asia and Oceania, Leiden University.

Reesink, Ger P. 1999. *A Grammar of Hatam. Bird's Head Peninsula, Irian Jaya*. Canberra: Pacific Linguistics.

Reesink, Ger P. 2002. "A Grammar Sketch of Sougb." In *Languages of the Eastern Bird's Head*, ed. by Ger P. Reesink, 182–276. Canberra: Pacific Linguistics.

Ricaut, François-Xavier, Nicolas Brucato, and Murray P. Cox. 2024. "Human Genetic Approaches to Malayo-Polynesian Prehistory." In *The Oxford Guide to the Malayo-Polynesian Languages of Southeast Asia*, edited by Alexander Adelaar, and Antoinette Schapper, 35–48. Oxford: Oxford University Press.

Riesberg, Sonja. To Appear. "Documenting and Describing the Languages of Eastern Indonesia — Past, Present, and Future." In *Indonesian Languages and Linguistics: State of the Field*, ed. by Abigail C. Cohn, Maya R. Abtahian, J. Joseph Errington, Thomas J. Conners, and Yanti. Ithaca: Cornell University Press.

Riesberg, Sonja. 2021. "Introduction to the Yali — English — German Dictionary With a Short Grammatical Sketch." In *A Yali (Angguruk) — English — German Dictionary*, ed. Sonja Riesberg, 1–49. Canberra: Asia-Pacific Linguistics.

Ross, Malcolm. 2017. "Languages of the New Guinea Region." In *The Cambridge Handbook of Areal Linguistics*, ed. by Raymond Hickey, 758–820. Cambridge: Cambridge University Press.

Saragih, Chrisma F. 2012. The Practical Use of Person Reference in Papuan Malay. M.A. Thesis, Radboud University, Nijmegen.

Samaun. 1979. The system of the contracted forms of the vernacular bahasa Indonesia in Jayapura, Irian Jaya: A project presented to the SEAMEO Regional Languages Centre. Singapore: SEAMEO Regional Language Centre.

Sawaki, Christian P. R. 2022. Language Maintenance and Language Perception of the Doreri Speaking People. B.A. Thesis, Universitas Papua.

Sawaki, Yusuf. 2005. *Melayu Papua: Tong pu bahasa*. Unpublished Manuscript.

Sawaki, Yusuf. 2007. "Does Passive Exist in Melayu Papua?" *Presentation, 11th International Symposium on Malay/Indonesian Linguistics (ISMIL 11)*, Universitas Papua, Manokwari, 6–8 August, 2007.

Sawaki, Yusuf. 2016a. A Grammar of Wooi: An Austronesian Language of Yapen Island, Western New Guinea. Doctoral Dissertation, Australian National University.

Sawaki, Yusuf. 2016b. "Teluk Wondama: Bahasa dan Masyarakat. Melacak Relasi dan Interaksi Sosial melalui Aspek Kebahasaan". *Working Paper* Seri 3, Manokwari: Universitas Papua.

Sawaki, Yusuf, Jeanete Lekeneny, Anna Rumakeuw, and Sonja Riesberg. 2016. "Language Documentation and Capacity Building in West Papua: The Center for Endangered Languages Documentation, Universitas Papua." Presentation, Konferensi Internasional Masyarakat Linguistik Indonesia (KILMLI), Udayana, Denpasar, 23–27 August, 2016.

Sawaki, Yusuf. 2018. "Meneropong Tipologi Bahasa-Bahasa di Papua: Suatu Tinjauan Singkat." *Linguistik Indonesia*, 36 (2): 129–143.

Sawaki, Yusuf, and I Wayan Arka. 2018. "Reflections on Linguistic Fieldwork and Language Documentation in Eastern Indonesia." *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 256–266. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24827

Sawaki, Yusuf, and I Wayan Arka. To Appear. "The Contemporary Sociolinguistics of Western New Guinea." In *The Oxford Guide to the Papuan Languages*, ed. by Nicholas Evans, and Sebastian Fedden. Oxford: Oxford University Press.

Scott, Graham R., Hyun Kim, Benny E. W. Rumaropen, Eleonora L. Scott, Christian G. Nussy, Anita C. M. Yumbi, and Robert C. Cochran. 2008. *Tong pu bahasa: A Preliminary Report on Some Linguistic and Sociolinguistic Features of Papuan Malay*. Sentani: SIL International Indonesia.

Silzer, Peter James. 1979. *Notes on Irianese Indonesian*. Jayapura: SIL & Universitas Cenderawasih.

Silzer, Peter James. 1983. Ambai: An Austronesian Language of Irian Jaya, Indonesia. Doctoral Dissertation, Australian National University.

Spriggs, Matthew. 1996. "Chronology and Colonization in Island Southeast Asia and the Pacific: New Data and an Evaluation." In *Oceanic Culture History: Essays in Honour of Roger Green*, ed. by Janet Davidson, Geoffrey Irwin, Foss Leach, Andrew Pawley, and Dorothy Brown, 33–50. Dunedin: New Zealand Archaeological Association.

Stasch, Rupert. 2013. "The Poetics of Village Space When Villages are New: Settlement Form as History-Making in West Papua." *American Ethnologist*, 40 (3): 555–570. https://www.jstor.org/stable/24027527.

Timmer, Jaap, and Chris Ballard. 2003. *An Annotated Bibliography for the communities of Berau Gulf and Bintuni Bay, Papua (Irian Jaya), Indonesia*. Research School of Pacific and Asian Studies, Australian National University, Canberra.

van den Heuvel, Wilco. 2006. Biak: Description of an Austronesian Language of Papua. Doctoral Dissertation, Vrije Universiteit, Amsterdam.

Visser, Eline. 2022. *A Grammar of Kalamang*. Berlin: Language Science Press.

Woodbury, Tony. 2003. "Defining Documentary Linguistics." *Language Documentation and Description* 1: 35–51. London: SOAS.

Woodbury, Tony. 2011. "Language Documentation." In *The Cambridge Handbook of Endangered Languages*, ed. by Peter K. Austin, and Julia Sallabank, 159–186. Cambridge: Cambridge University Press.

Yoder, Brendon. 2018. "The Abawiri Tone System in Typological Perspective." *Language*, 94 (4): e266–e292.

CHAPTER 4

Diversifying science

Insights of native- and non-native-speaker collaborations in linguistic descriptions

I Wayan Arka
Australian National University | Universitas Udayana

This chapter reflects on over two decades of linguistic research in Indonesia, examining collaborations between native- and non-native-speaker linguists and local speech communities in documenting and describing minority and endangered languages. Drawing on case studies from Balinese, Rongga, Marori, and Enggano, it explores how positionality, authority, and co-production shape grammar writing and broader documentation outcomes. Situating these experiences within the shift toward community-led, ethically grounded, and ethnographically informed fieldwork, the chapter underscores the importance of capacity building for native-speaker linguists and sustained engagement with both communities and local institutions. It argues that partnerships grounded in mutual respect and shared epistemic goals can produce richer, more culturally embedded, and socially impactful linguistic descriptions, advancing a more inclusive and decolonial linguistic science.

Keywords: native-speaker linguists, collaborative grammar writing, positionality, authority and co-production, community-led documentation, ethically grounded fieldwork, capacity building, decolonial linguistics, endangered languages, sustained institutional engagement

1. Introduction

This chapter offers a critical reflection on my long-standing engagement with linguistic research in Indonesia, focusing on the evolving collaborative dynamics between native- and non-native-speaker linguists as well as speech communities in the documentation and description of languages. Indonesia's extraordinary linguistic diversity, with hundreds of Austronesian and Papuan languages, provides both a rich empirical base and a complex sociocultural terrain in which

https://doi.org/10.1075/slcs.240.04ark
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

issues of positionality, authority, and co-production are continually negotiated. In this context, building on Ameka's (2006) co-creative grammar writing, I reflect on and explore the recent paradigm shift in field linguistics (Bischoff & Jany 2018; Czaykowska-Higgins 2009; Dobrin & Berson 2011; Florey & Himmelmann 2010: 122; among others), discussing how partnerships between native speakers from local speech communities and non-native linguists — when grounded in mutual respect and shared epistemic goals — can lead to more culturally embedded, ethically sound, and analytically robust linguistic work, extending beyond grammar writing. This aligns with calls to reimagine linguistics as a decolonial, community-led practice that actively challenges the extractive legacies of the field (Hudley, Mallinson and Buchholtz 2024; Tsikewa 2021).

My reflections are based on over two decades of fieldwork and academic practice, drawing from three intersecting roles that have shaped my perspective. First, as a native-speaker linguist working on Balinese and Indonesian, I have experienced the advantages of insider knowledge — along with its limitations — and the transformative impact of collaborative analysis with non-native linguists. Second, as a non-native fieldworker documenting Indonesia's endangered and minority languages such as Rongga, Marori and Enggano. I have navigated the challenges of outsider positionality and developed models for collaborative engagement that centre the agency of local communities. Third, as a supervisor and mentor, I have supported numerous native-speaker students in documenting their own languages, guiding them through different challenges at various stages of their academic trajectories, and witnessing how their training and empowerment reshape the research landscape from within.

This chapter explores the affordances and tensions across these roles, with the aim of identifying practices that support inclusive, pluralistic, and co-productive linguistic science. Drawing on concrete case studies — from working on my native language Balinese to leading documentation projects in Tanah Papua and the Barrier Islands regions — I show how linguistic collaboration can thrive when all actors are supported in developing their capacities, and when community goals are integrated with academic objectives. In this view, collaboration is not simply a methodological convenience but an ethical and epistemological commitment.

However, as I have learned through extensive field experience (Arka 2010, 2013, 2015, 2025; Hisa et al. 2017, 2018), these collaborations only succeed when linguists' research agendas are developed alongside — and not at the expense of — the interests of the speech communities themselves. As Dobrin and Schwartz (2016) emphasize, collaboration is not a universally appropriate or sufficient model for every context. In many minority settings, linguistic work intersects with broader community aspirations around cultural revitalisation, education, identity, and economic-cultural-social wellbeing. As such, linguistic documentation must be

framed not only as a scholarly enterprise but also as a socially responsible practice. This has become a central philosophy in my work: that research is most meaningful when it contributes reciprocally to the communities it engages with, and when it is co-constructed by those who speak, live, and sustain the languages we seek to describe.

In what follows, I outline the conceptual grounding for this co-creative approach by, first of all, discussing the what and how of linguistic research and language description (Section 2). I then examine the shifting roles of native- and non-native-speaker linguists (Section 3), and reflect on challenges and strategies for equitable collaboration (Section 4). Along the way, I highlight insights gained through supervising native-speaker students (Section 5), present case studies — Balinese and Enggano — that illustrate co-creation in practice (Section 6), and formulate recommendations for the future of linguistic science (Section 7), with the broader goal of diversifying the scientific enterprise of linguistics through more inclusive and locally rooted forms of knowledge-making. Conclusions are then given in Section 8.

2. On the what and how of linguistic descriptions and language documentation

A fundamental challenge in linguistic description and language documentation, especially when writing a grammar, lies in determining *what* to document and describe: which phenomena to prioritise, what data to collect, and what kinds of knowledge to foreground. This task is far from straightforward. Grammar writing, as Evans and Dench (2006:1) remind us, "is a demanding art which calls for the integration of form and meaning, system and use, synchrony and diachrony." It becomes even more complex in the context of language documentation, particularly for minority and endangered languages where linguistic competence among speakers may be uneven, and sociocultural factors heavily shape the ecology of language use. My own fieldwork experience across Indonesia has shown that producing a grammar is only one piece of a much larger puzzle. Alongside what Evans and Dench (2006) call the "grammar-dictionary-text" trinity, — or what Musgrave and Thieberger (2021) term the "documentation quartet," with the addition of media recordings in modern documentary linguistics — there are local expectations, needs, and constraints, including the demand for hypertext and interlinking in digital materials. It is not enough to produce grammars as scholarly artefacts; our work must also respond to what I have elsewhere called the "multi-dimensionality of capacity building" (Arka 2025), which includes fostering local participation, leadership, and community relevance. In addition, in today's

digital age, it is increasingly valuable to leverage hypertext and web technologies to enhance accessibility and verifiability of analytical claims — for example, by allowing users to view examples in full context and access supporting evidence.

In thinking about what kinds of linguistic descriptions we produce, I align with Evans and Dench's call to recognise the interdependence of descriptive, typological, and formal-theoretical approaches. They describe these as "a triadic and mutually complementary relationship" (2006: 2). Descriptive linguistics provides grounded accounts of particular languages, typology brings out patterns across languages, and formal linguistics constructs predictive models. In my own documentation projects — whether on Rongga, Marori, or Enggano — I have found it crucial to design the descriptive work so that it not only serves the immediate needs of language maintenance (e.g. orthography development, or school materials), but also informs theoretical questions (e.g. on alignment, voice, or clause structure) and typological debates (e.g. on Austronesian variation). Descriptive work, as I have argued (Arka 2025: 32), is the logical entry point before comparative or formal analysis can be carried out, but it must be embedded in a larger strategy that bridges academic inquiry and community benefit.

This multi-stranded goal also requires expanding the very idea of what constitutes a "grammar". As Evans and Dench emphasise, "a grammar is inescapably an interpretive account" (2006: 7), shaped by the linguist's choices, theoretical assumptions, and intended readership. From my perspective, working as both a native and non-native researcher in different language contexts, I find that such interpretive work must also be shaped by the language community's aspirations, the available human and institutional resources, and the evolving dynamics of identity and motivation. In an ethnosyntactic or ethnogrammatical sense (Enfield 2002), these priorities influence the content and organisation of the grammar: which categories are described in depth, what examples and varieties are chosen, and how explanations are framed. These considerations are especially critical when grammar writing also serves pedagogical purposes, such as developing teaching materials for local schools. In Enggano, while the actual drafting of the grammar has been primarily undertaken by the non-local linguist members of our project team, the process has been co-creative and evolving, informed by ongoing community consultation through our Community Research Collaborator (CRC). This work has proceeded alongside training the CRC in basic linguistics, ELAN, building a lexicon for school use, and developing literacy resources. In Marori, the descriptive work seeded a broader eco-cultural revitalisation project. These experiences underscore that in the contexts where I work, a grammar is not simply a product of analytical rigour — it is part of a broader co-creative enterprise involving multiple stakeholders, epistemologies, and goals. Descriptive linguistics, when situated in such multilingual, dynamic, and politically charged

environments, is both a scientific and a deeply social act. Grammar writing, at different levels, is thus a co-constructed account in which analytical decisions are shaped and informed by community knowledge, values, and goals.

Understanding *how* to undertake this descriptive task requires equal methodological care. In recent years, linguistic fieldwork has undergone a shift — sometimes referred to as the "*documentarist turn*" — from decontextualised elicitation toward ethnographically informed, corpus-based, and participatory approaches (Himmelmann 1998; Woodbury 2003, 2011). This shift owes much to the early engagement of Nikolaus P. Himmelmann in the 1990s, whose advocacy for richly contextualised documentation laid a foundation for modern best practices, as well as to other early pioneers such as Anthony C. Woodbury. My own methodological stance builds directly on these foundational ideas, extending them in ways that respond to contemporary challenges of capacity building, community co-authorship, and ethical reciprocity in field linguistics in Indonesia. Traditional elicitation methods — such as sentence translation — often fail to capture real-time interactional dynamics like register alternation, politeness strategies, or epistemic stance. To address this, modern documentation increasingly integrates three key practices: the structured collection of metadata, the development of multimodal corpora, and ethnographic grounding. Recording demographic, genealogical, and contextual metadata about speakers enables researchers to link linguistic variation to socially meaningful variables. Corpus design must include naturalistic and stimulus-based materials that are sensitive to pragmatic domains often underrepresented in spontaneous data. And crucially, these methods must be informed by participant observation and community engagement. This ethnographic lens ensures that linguistic patterns are not misinterpreted, and that grammars reflect the lived sociocultural realities of speakers. As we have argued (Arka & Mullan, to appear), these methodological shifts are essential, for instance, in capturing how language encodes social cognition, and they allow our grammatical descriptions to speak not just to formal and typological concerns, but also to the broader sociocultural contexts in which language is embedded.

In navigating these complex demands of grammar writing (or language descriptions more generally) serving typological generalisation, formal rigor, and community relevance, I have found that the choice of theoretical framework plays a critical role. It is essential to work with a framework that is not only sufficiently modular and flexible to accommodate the unique structural and sociocognitive patterns of individual languages, but is also capable of revealing cross-linguistic parallels in a principled way. Lexical-Functional Grammar (LFG) offers such a framework (cf. Kroeger 2007). Its parallel and modular architecture enables analysts to tease apart and explicitly model multiple interacting layers of linguistic structure, including constituent structure, grammatical relations, information

structure, and social meaning. As we argue in Arka and Mullan (to appear), LFG "provides an effective framework for analysing the intersection of social cognition and typological variation," particularly through its capacity to model context-sensitive grammatical behaviour without flattening language-specific distinctions.

This is especially valuable in multilingual, socially stratified, and contact-heavy ecologies like those in Indonesia, where social cognition is deeply woven into grammar. The discourse pragmatic structure (prag-str) in LFG, for instance, allows us to explicitly encode culturally specific information (e.g. social caste-based relations, *menak* 'high(er) caste' and *jaba 'commoner'*) that determines lexical and grammatical choices, such as in Balinese speech levels or polite pronominal forms. This linguistically relevant information analysed as "social predicate" can be linked to grammatical modules like f-structure (functional structure) and a-structure (argument structure) through clearly defined correspondences. Such a modular setup lets us account for how variation in language use is shaped by speaker — addressee dynamics, social hierarchy, politeness, and voice constructions — while keeping these components separable for different analytical purposes. This modularity also makes LFG collaborative-friendly: different researchers, even from divergent theoretical traditions, can contribute to the same description by working on distinct but compatible layers (e.g. one on syntax, another on discourse, another on semantics) without requiring full adoption of LFG's formal machinery. In my own work, this flexibility has enabled productive partnerships in grammar writing, both with fellow LFG specialists and with typologists and documentary linguists who use other frameworks. Crucially, this blend of language-specific descriptive adequacy and cross-linguistic comparability, and theoretical openness makes LFG especially well-suited for the kinds of collaborative grammar work I undertake — whether as a native-speaker linguist working on Balinese and Indonesian or a non-native researcher documenting languages like Marori and Enggano.

3. Native-speaker and speech communities in linguistic description and documentation

My early linguistic training and later academic development have been shaped by a bilingual upbringing in Balinese and Indonesian. As a native-speaker linguist, I bring to my analyses an intuitive grasp of sociocultural nuances and communicative practices that often elude external analysts. This insider position enables more in-depth interpretations of subtle pragmatics and lexicogrammatical features, offering unique insights into language description and into specific areas prioritised during documentation. However, the benefits of this insider role can only be fully realised — and its potential pitfalls mitigated — through quality linguistic

training and sustained collaborative research experience. Such pitfalls include taking implicit knowledge for granted or overlooking deeper structural patterns due to over-familiarity, even among trained linguists (Section 3.2). Before discussing the importance of linguistic training for native speakers, I first examine the diverse roles that native speakers can play in linguistic description (Section 3.1).

3.1 Diverse roles and engagements

As a linguist working across multiple roles — both as a native speaker describing my own languages and as a non-native researcher documenting others — I have come to appreciate that linguistic description is most meaningful when conceived as a collaborative, socially grounded enterprise. Native speakers and local communities contribute to our research in a variety of ways, each role offering different kinds of insights and support. At one end of the spectrum are relatively passive language consultants who participate by providing elicited responses or contributing to recordings of natural speech. Their involvement is often shaped by the researcher's agenda, and while crucial, it may be limited to the provision of linguistic forms rather than interpretive or contextual information.

More engaged are those whom I consider relatively active language consultants — often local elders, teachers, or community leaders with a deep understanding of their language and its sociocultural foundations. In my experience, this group also includes younger, well-educated native speakers who, despite lacking formal linguistic training, offer informed perspectives on what linguistic or cultural domains are important to document. Their input has guided decisions on documentation priorities, corpus planning, and even methodological choices. These collaborators play an important role in shaping the research trajectory, serving not just as data sources but as cultural interlocutors and knowledge custodians. This aligns with Ameka's (2006:70–71) call for native speakers to be seen as agents who "help shape the record" and whose contributions are essential to generating "optimal descriptions".

A third and especially vital role is that of the local research assistant, or Community Research Collaborator (CRC) — an individual who not only functions as a language consultant, combining aspects of both passive and active involvement, but also serves as a liaison and coordinator, managing field logistics, building trust within the community, and facilitating interactions between the external researcher and the broader speech community. The importance of this role cannot be overstated. In my fieldwork across Indonesia, the success and sustainability of language documentation efforts have often depended on the recruitment and training of at least one committed and capable local CRC. Such individuals become the anchor of long-term, locally embedded linguistic research, ensur-

ing that the project's impact extends beyond the typical 2–3 year lifespan of an outsider's funded initiative. They help institutionalise documentation processes, contribute to the community's linguistic and cultural resilience, incorporate indigenous epistemologies and research methodologies, as well as support the emergence of future native-speaker linguists — thus embodying the kind of co-creative model advocated by Ameka (2006: 95) and Tsikewa (2021), also echoed in my own work (Arka 2025).

In sum, meaningful linguistic documentation must be attuned to the diverse roles that native speakers and local communities can play. Each form of involvement, from the quiet contribution of a consultant in a recording session to the strategic guidance of a local CRC, enriches the descriptive record and strengthens its relevance for both academic and community goals. Our task as linguists is to recognise, respect, and cultivate these roles, ensuring that our work is grounded in reciprocal relationships and shared epistemic authority.

3.2 On the need of native speaker training in linguistics

The growing recognition of native speakers as epistemic collaborators in linguistic research must be accompanied by a serious commitment to training. While native speakers bring irreplaceable cultural knowledge, lived experience, and intuitive linguistic insight to the table, they often lack access to the analytical tools, methodological rigour, and theoretical frameworks that structure academic linguistics. As Felix Ameka (2006: 75) notes, "the most powerful research machine imaginable is the trained linguist who is also a native speaker of the object language." Yet he also reminds us that being a native speaker linguist is a "mixed blessing" (2006: 76): while the depth of intuitive knowledge may yield "flashes of insight," such insight can be compromised by theoretical bias or a narrow view shaped by the speaker's own idiolect or social positioning.

There are a growing number of native speakers of Indonesian languages (with myself as an example) who are committed to serious training in linguistic theory and method. As I have argued with Mullan (Arka & Mullan, to appear), grammars are not neutral inventories of structure — they are interpretive models that benefit from both linguistic theory and sensitivity to speaker-based social cognition. The trained native speaker, especially when equipped with an understanding of grammatical architectures like LFG, can bring to light subtle yet systematic dimensions of meaning encoded in everyday language use, such as sociopragmatic distinctions and interactional motivations. These are frequently obscured in purely formal or typologically driven accounts.

In my own practice — working as a native-speaker linguist on Balinese and Indonesian, and as a non-native researcher on languages like Marori and Enggano —

I have seen first-hand how training native speakers in basic linguistics transforms the documentation enterprise. The basic training in linguistics for local RAs enables them not only to contribute as informed consultants, but to take on active roles in data collection, transcription, analysis, and publication. It also creates a more equitable collaborative space, where descriptive and analytical decisions are co-owned. This training need not always take place within the traditional university setting. In the projects that I have led, we have pursued context-sensitive, field-based training — through hands-on workshops on ELAN and Praat, guided corpus annotation, collaborative text glossing, and participatory grammar-writing sessions. These activities build confidence, competence, and a sense of ownership, while demystifying the analytical process.

As for advanced training in linguistics, based on both my own academic training and extensive supervisory experience, I have seen that such training in linguistics — at undergraduate and postgraduate levels — enables native speakers to contribute meaningfully across multiple stages of research, significantly improving the efficiency of documentation and the depth and quality of analysis. This is particularly evident in my supervision of PhD students who are native speakers working on their own languages: their insider knowledge, combined with formal linguistic training, has led to richly informed and theoretically grounded descriptive work. However, such outcomes are only possible with sustained and targeted support, as advanced training — especially at the PhD level — is time-intensive, costly, and often takes several years to yield tangible results. Moreover, it is a major personal and academic undertaking, and in many communities, there may be few individuals who have both the capacity and willingness to commit to such a demanding path.

In general, through a combination of context-sensitive, field-based mentoring and structured analytic exposure, I have seen native speaker collaborators grow from data providers into confident co-analysts. For example, the development of the *pragmatic structure* (prag-str) module within LFG (Arka & Mullan, to appear, Section 5.2) owes much to my native-speaker intuitions and life-experience about culturally salient predicates — like the Balinese concepts of *menak* and *jaba* — which map directly onto syntactic and information-structural behaviour. Likewise, the multilingual Enggano dictionary (Rajeg et al. 2025) is an outcome of years of collaboration with local Enggano native speaker experts. Without a native and shared analytic vocabulary and metalinguistic awareness, these intuitions might never have found formal expression in grammar and dictionary.

For long-term benefits, training native speakers must go beyond technical skills in linguistics, and ideally include some training in related disciplines and leadership skills, with the hope that this will cultivate their capacity to work on their language and culture through broader interdisciplinary perspectives with

practical benefits (e.g. standardisation and needs analysis for teaching material development and local literacy, as well as initiating/managing language maintenance or revitalisation programs). It should include critical reflection on linguistic ideology, dialect diversity, and the politics of representation: Whose variety is being described and used as a "standard" to be taught at local schools? Who defines what counts as "core grammar"? How do social categories like gender, caste, or kinship inform grammatical descriptions and their use? Is there a need to establish a local language committee to oversee the development of textbooks and a dictionary? Who decides membership and leadership in such a committee? As Ameka (2006: 83–84) cautions, native speaker linguists can also fall into the trap of prescriptive bias or fail to reflect on the social variation within their language. This risk may be mitigated in the modern Indonesian context, where internet access is now available even in remote areas, enabling community members to access rich, interlinked resources as described by Musgrave and Thieberger (2021). Such community access and engagement allow users to move seamlessly among resources and to become aware of the linguistic variability in their language, which can then be included in examples, dictionary entries, and teaching materials. Thus, robust training programs must engage students with theoretical pluralism, sociolinguistic sensitivity, and an awareness of the historical, political, and ecological dimensions of language work. This aligns with our argument (cf. Arka 2025; Arka & Mullan, to appear) that linguistic analysis must remain accountable to the community's perspectives and communicative norms, not just to the demands of typology or theory.

Ultimately, training native speakers in linguistics is not only about producing better grammars. It is about restructuring the ecology of linguistic science itself — expanding who gets to speak authoritatively about language, and what kinds of knowledge count. It is also key to sustaining long-term engagement: many of my most successful documentation projects owe their longevity not to the periodic visits of outside researchers, but to the persistence and leadership of trained local collaborators, including community-based research assistants who have grown into independent researchers. These individuals do not merely extend the reach of external researchers; they become institutional bridges and cultural stewards, managing corpora, coordinating fieldwork, and serving as mediators between local language ideologies and academic research goals. In short, training native speakers in linguistics is not only a methodological necessity — it is an ethical and epistemological investment in a more inclusive and co-creative model of linguistic science.

4. Non-native linguists and (minority) Indonesian languages

While the previous section focused on the contributions, capacities, and evolving roles of native-speaker linguists and community members in documentation, this section shifts perspective to my work as a non-native linguist documenting minority endangered languages across Indonesia. In this role, I am an outsider navigating unfamiliar linguistic ecologies and complex sociocultural terrains. The challenges in this role are considerable, ranging from the need to build trust with community members to managing expectations (particularly around financial compensation), and contending with residual suspicions associated with the term *proyek* in the Indonesian context. In this subsection, I first provide some historical background in Section 4.1, outlining the foundational contributions of non-native linguists — primarily missionaries — during both the pre- and post-Indonesian independence periods. In Section 4.2, I examine the challenges encountered by non-local linguists, whether Indonesian or foreign, in undertaking documentary research within present-day Indonesia.

4.1 Contribution of non-native linguists in Indonesia: A historical overview

In reflecting on the history of linguistic research in Indonesia, I acknowledge that much of the foundational work on both dominant and minority languages has been carried out by non-native linguists, particularly foreign researchers, since the Dutch colonial period. These early efforts, beginning in the 19th century, were led primarily by Christian missionaries who had the training, resources, and institutional support that local communities at the time could not access. Figures such as *H.N. van der Tuuk*, who compiled a grammar and bilingual dictionary of Batak, and *J.A. Verheijen*, who worked extensively on Manggarai — including ethnobiological and ethnographic studies — set the stage for future linguistic work. Similarly, *Petrus Drabbe* contributed valuable grammatical descriptions and texts on the languages of Tanimbar and southern New Guinea. These early works — though shaped by religious and colonial motivations — provided a lasting record that continues to inform current linguistic and typological research (Arka 2025: 5).

This tradition of non-native linguistic engagement continued into the post-independence era. While Indonesia gained political autonomy in 1945, the linguistic research landscape remained heavily influenced by foreign institutions and researchers. In the second half of the 20th century, organisations such as the Summer Institute of Linguistics (SIL) played a major role in documenting local languages across the archipelago, particularly in remote regions. In some areas, SIL's presence lasted for decades before their activities were restricted due to political

and religious sensitivities (Arka 2025: 6–7). Their contributions, however, were undeniably significant — not only in producing descriptive materials, but also in introducing methods of data collection and analysis that shaped future generations of linguists, including my own.

Since the turn of the 21st century, and particularly following the *documentarist turn* (cf. Himmelmann 1998; Woodbury 2003), a new phase of non-native linguists' engagement in Indonesia has emerged — one shaped by modern, corpus-based, community-engaged language documentation. A milestone in this development, especially in Tanah Papua, was the establishment of the Center for Endangered Languages Documentation (CELD) at the University of Papua in Manokwari. This centre was launched alongside the Wooi Documentation Project, funded by DoBeS (*Dokumentation Bedrohter Sprachen*), under the supervision of Nikolaus Himmelmann. More than a single project, this initiative set a precedent for building both documentary corpora and local capacity. CELD went on to coordinate major documentation efforts, including the "highlands summit" program in 2012 (targeting Yali, Walak, Western Dani, Eipo, and Yale), the 2014 projects on Iha and Mpur, and the 2019–2022 ELDP-funded documentation of Konda (Yaben) and Kais.

Himmelmann's role was pivotal not only in supervising individual projects but also in fostering a national network of trained Indonesian linguists. His leadership extended to capacity-building initiatives such as the landmark national language documentation workshops held in Bali in 2006–2007 (Florey & Himmelmann 2010, cf. also Mayani & Katubi, this volume), in which I participated as an instructor. This model — bringing together participants from across Indonesia, including both minority and non-minority language speakers — proved transformative: several participants went on to complete PhDs and become leaders in their own right, among them was Yusuf Sawaki, now Director of CELD. The ripple effects of this model can still be seen today in similar workshops run by other institutions, including the *Pusat Riset Preservasi Bahasa dan Sastra* (Centre for the Preservation of Language and Literature) and BRIN (National Agency for Research and Innovation). These workshops have combined technical training in tools such as ELAN and FLEx with guidance on grant writing and project design, a critical combination for enabling Indonesian linguists to secure competitive funding and lead independent documentation projects on the international stage.

At the same time in more recent years, I have seen a gradual but important shift: more Indonesian linguists — especially those working for Badan Bahasa or as academics at universities — have taken up the task of documenting and describing the languages of the archipelago. However, these researchers are often still non-native to the languages they study, especially in the case of minority languages spoken in remote or marginalised regions. This reflects a persistent structural limitation: the communities where many of these languages are spoken often

lack access to higher education and have limited pathways for their members to become trained linguists. In many of the places where I have worked — such as Merauke, central Flores, and the Barrier Islands — there simply were no local individuals from the minority speech communities with the linguistic training and experience necessary to lead research projects.[1]

Because of this, modern language documentation in Indonesia — including many of my own projects — has continued to depend on external support, often through international collaborations funded by bodies such as ELDP, DoBeS, and the AHRC. These collaborations have resulted in rich, publicly available multimedia corpora and have contributed to typological and theoretical inquiry, as well as to language maintenance. Yet they also highlight the ongoing need for capacity building and the continued relevance of non-native linguists in supporting this work (Arka 2007, 2025; Florey & Himmelmann 2010). From a historical and practical standpoint, the role of non-native linguists remains central — not as replacements for local leadership, but as collaborators and facilitators who can help fill critical gaps while working towards a more inclusive future for linguistic research in Indonesia.

1. In recent decades, an increasing number of Indonesian linguists — including those from non-minority speech communities such as Nias — have completed advanced degrees (MA and PhD) in linguistics, particularly from Indonesian universities. This growing cohort represents an important potential force for advancing language documentation within Indonesia. However, based on my experience serving on the international ELDP panel reviewing grant applications from Indonesia and Southeast Asia more broadly, it is evident that many of these emerging scholars still require greater exposure to collaborative research practices and training in the full range of documentation methodologies in order to compete effectively on the international stage. To date, only a small number of Indonesian linguists have successfully secured highly competitive international grants such as ELDP and ELF (Endangered Language Funds) funding independently — that is, without the partnership or mentorship of foreign collaborators. Recognizing this gap, initiatives were launched to support local capacity building, including the ELDP- and Volkswagen-sponsored language documentation workshops held in Indonesia (Florey & Himmelmann 2010), which, as mentioned above, inspired similar workshops organized by (local) Indonesian institution. Recently (2019–2025), the OCSEAN, an MSC RISE exchange program funded by the European Commission (https://www.ocsean.eu/), also includes capacity building for Indonesian scholars through hands-on training in data collection and analysis, developing transferable skills, and enabling community-led research with lasting impact.

4.2 Challenges of non-local linguists in linguistic and documentation research

While non-local linguists — especially foreign scholars — have made essential contributions to the documentation of minority languages in Indonesia, their role is not without complexity. As outsiders, they face a range of methodological, ethical, and practical challenges that shape the effectiveness, sustainability, and reception of their work. These challenges must be understood not as individual shortcomings, but as structural and relational dynamics embedded in sociocultural, educational, and historical contexts (Arka 2018; Sawaki & Arka 2018: 259–260).

4.2.1 *The role of the outsider*

Non-local linguists often operate with positionality shaped by their status as outsiders — linguistically, socially, and institutionally. In my experience, the success of documentation projects is often significantly influenced by how non-local researchers manage this position of difference. Gaining access to communities, building trust, and negotiating one's role often take considerable time and sensitivity. In remote and marginalised communities across eastern Indonesia, where linguistic diversity is high but educational and infrastructural support is low, researchers may be met with suspicion, indifference, or misplaced expectations — especially in places historically conditioned by missionary or government development projects (Sawaki & Arka 2018: 263–264).

This positionality also extends to practical fieldwork issues. In many areas, foreign linguists are better trained, resourced, and institutionally supported, making them more capable of initiating and sustaining documentation projects. However, this advantage can also deepen dependency and limit opportunities for local scholars and communities to assume leadership roles. As noted in Sawaki and Arka (2018: 260), "foreign linguists are better equipped for fieldwork… [and] are in a better position to generate publications, organize seminars, and facilitate training activities." This asymmetry underscores the importance of recognising and mitigating the implications of outsider-led documentation.

4.2.2 *Methodological and ethical challenges*

One of the most persistent challenges for non-local linguists is negotiating expectations. Communities often associate the term *proyek* 'project' with government-initiated programs that distribute funds with little accountability, leading to the assumption that any outsider project is a cash-disbursement activity. This has led to misunderstandings and frustration when documentation projects do not align with these expectations. As a strategy, I and colleagues in the Center for Endangered Languages Documentation (CELD), for example, have found it more effec-

tive to use the terms *program* or *aktivitas* 'activity' to manage expectations and communicate goals more clearly (Sawaki & Arka 2018: 263).

Ethical challenges also arise in ensuring that documentation is not extractive, highlighting a clear call for decolonial, community-engaged practice.[2, 3] These challenges echo the critiques raised by Tsikewa (2021) and Hudley, Montoya et al. (2024), who argue that dismantling these entrenched imbalances requires systemic reform in how linguistics is taught, funded, and institutionally rewarded. Misinterpretation of linguistic data due to unfamiliarity with local norms, culturally inappropriate recordings, or exclusion of relevant community voices can all compromise the integrity of the documentation. Sawaki and Arka (2018: 262) also report how local rivalries and clan-based tensions may complicate the recruitment of consultants or the sharing of cultural knowledge. Furthermore, some community members may engage in projects primarily for financial compensation, contributing data that lacks contextual or narrative depth. Together, these challenges point to the need for dialogue-based, community-participatory methods that foreground reciprocal understanding and collective goals. In any case, when navigating challenging local dynamics (e.g. internal community divisions, shifting allegiances or unpredictable politics; cf. Dobrin & Schwartz 2016), it is wise to remain flexible and guided by ongoing reflexivity and sensitivity.

4.2.3 *Learning through ethnographic engagement*

Despite these challenges, non-local linguists can contribute meaningfully when they engage in ethnographic listening and long-term collaboration — particularly to gain a deeper understanding of the social meanings behind linguistic choices (Briggs 1986: 3–5), especially in contexts of multilingualism and identity negotiation (Dobrin & Berson 2011). Many valuable insights emerge not from formal elicitation sessions but from participating in rituals, attending communal events, or sitting with elders who share linguistic and ecological knowledge — the very goals

2. The move away from extractive practices in Indonesia, as in many other post-colonial contexts, reflects a broader reaction against the exploitative resource extraction characteristic of European colonial rule. In linguistics, this shift aligns with the wider agenda of decolonizing the field. In contemporary Indonesia, it also entails resisting extractive approaches by central authorities or Jakarta-based institutions toward local communities, echoing broader debates on local and regional autonomy (Arka 2007).

3. However, it should be noted that linguistic fieldwork or documentary linguistics exists on a continuum from traditional, linguist-focused models — where community members are mainly sources of data — to fully collaborative, community-based models that centre community agency, expertise, and priorities (Czaykowska-Higgins 2009). Methodologically, research relationships must be tailored to the specific goals, needs, and aspirations of each community; what is collaborative and appropriate in one context may not be so in another.

of language documentation and documentary linguistics (cf. Himmelmann 1998; Woodbury 2003, 2011). For example, in our ethnobiological work with Marori and Wooi speakers, we documented not only the language itself but also how it encoded social-ecological knowledge about sago and mangrove ecosystems in traditional and contemporary contexts — insights that could only be accessed through culturally immersive fieldwork (Hisa, Mahuze & Arka 2017: 258; Sawaki & Arka 2018).

A key lesson is the importance of working with and through local leaders and intermediaries — individuals who understand both community values and research goals. Involving respected local figures early on in the project and inviting them to help shape participation strategies mitigates conflicts and enhances buy-in (Sawaki & Arka 2018: 263–264). Moreover, when trained locals are empowered to carry out documentation themselves, the work becomes more sustainable and more deeply rooted in community perspectives.

5. Supervising Indonesian native-speaker students: A transformative process

One of the most promising developments in Indonesian linguistics in recent decades is the growing number of native-speaker students engaging in the documentation and analysis of their own (heritage) languages. I use *heritage* here in the sense of Polinsky and Kagan (2007: 369) and Montrul (2016: 16) — that is, languages acquired in the home during childhood but which, due to a later shift in language dominance, are no longer the individual's strongest or primary language. Some of these students could be described as *semi-native* speakers: while they possess valuable insider cultural knowledge and at least partial competence in the language spoken by their parents, they are now dominant in Indonesian, the national language, whether because they moved with their families to urban centres such as Java or because Indonesian became the lingua franca among their peer group even within their home communities. In such a situation, full fluency is often confined to the older generations. When equipped with robust analytical tools and theoretical frameworks, such students are uniquely positioned to contribute to a form of scholarship that bridges local epistemologies and academic rigour in their research. In my supervisory role over the past two decades — primarily through Udayana University and the Australian National University (ANU) — I have found the supervision and long-term mentoring of these students to be a deeply transformative process, both for the students and for myself as a supervisor.

The needs of native-speaker students vary depending on whether they pursue graduate education domestically in Indonesia or overseas. For those studying abroad, particularly in English-speaking countries with high-ranking institutions (e.g. Australia, the UK, and the US), additional preparation is often required.[4] This includes securing competitive scholarships, meeting English proficiency standards, and adjusting to different academic cultures and expectations. This preparatory stage, based on my experiences helping Indonesian students be accepted at the ANU, takes years. In contrast, these requirements are often less pressing for students undertaking postgraduate studies domestically. However, even within Indonesia, the rising emphasis on international academic standards — such as the requirement to publish in peer-reviewed journals for thesis completion — has increased pressure on students and, by extension, on supervisors, to provide more structured academic support.

In today's globalised academic landscape, students are increasingly expected to demonstrate scholarly productivity and competitiveness on the international stage, even for local academic and research positions in Indonesia. While top-ranked international universities may not formally require publications during doctoral study, publishing in reputable journals has become common practice as it signals a strong academic trajectory and enhances employability in a competitive job market. For many native-speaker students, particularly first-generation scholars, guidance in navigating the publication process is essential and becomes a core part of the supervisory relationship.

From a supervisor's perspective, supporting PhD students — especially those working on their own languages — has been a rewarding and intellectually generative experience. My large research projects have always been conceived with local capacity building in mind; hence involving Indonesian PhD students (e.g. the *Barrier Islands Project*), or Indonesian Postdoctoral fellows (e.g. the *Enggano Project*) in the team. These Indonesian scholars are not only pursuing independent PhD goals, but are also contributing meaningfully to collaborative, team-based

4. While this discussion focuses on English-speaking countries because English proficiency is a universal requirement for publication in leading international journals (Cargill & Burgess 2017: 3–4), the point applies more broadly. In countries where the medium of instruction is not English, students generally need English in addition to the national language. For example, in Germany or France, graduate students must typically demonstrate competence in the local language for everyday life and administrative purposes, while also maintaining a high level of academic English. In contrast, in some European contexts such as the Netherlands, many graduate programs — especially in the humanities and social sciences — are fully taught in English, so Dutch is not an admission requirement. Similarly, in Japan, while most universities require Japanese proficiency, a number of institutions offer postgraduate programs entirely in English, admitting students without prior Japanese language skills.

research with shared objectives. Their projects serve as integral components of broader research and documentation efforts, and they benefit from participating in collective activities such as joint fieldwork, data processing, analytic workshops, and co-authored publications.

Such collaborative models provide a dynamic research environment in which students are mentored to work both independently and as team players. They are trained across all stages of research — from planning and fieldwork to data analysis and dissemination. This experience equips them with critical academic and professional skills, preparing them to become local agents and future leaders in their fields, at least in Indonesia. From my standpoint as a supervisor, the benefits are equally significant: these collaborations have generated insights that would not have emerged through individual effort alone. Regular individual supervision and group meetings have fostered a vibrant intellectual exchange, enriching the research and strengthening the outcomes of both student and team-led work.

These experiences underscore the vital role that capacity building through graduate supervision plays in the future of linguistic scholarship in Indonesia. Supporting Indonesian native-speaker students through structured mentorship, international exposure, and collaborative research opportunities not only empowers individuals but also contributes to the broader goal of sustaining and revitalising Indonesia's rich linguistic heritage. By investing in this next generation of scholars — who are deeply rooted in their communities and increasingly connected to global research networks — we lay the groundwork for a more inclusive, resilient, and locally driven model of language documentation and analysis. This approach, I believe, is essential for ensuring the long-term impact of our collective work, both within academia and in the communities whose languages we seek to understand and support.

6. Co-creation in practice, case studies: Balinese, Marori and Enggano

In this section, I reflect on my collaborative experiences as both a native-speaker linguist and a non-native researcher, to illustrate how co-creation between native and non-native linguists can lead to richer, more nuanced, and culturally grounded language descriptions. Drawing on three case-study projects — Balinese, Marori and Enggano — I highlight how collaborative, interdisciplinary, and cross-positional approaches to language documentation and analysis have shaped my research practice over the years. The first case (Section 6.1) describes my experience as a native speaker of Balinese working with both Balinese and non-Balinese linguists. The second case (Section 6.2) turns to my role as a non-native researcher working with the native Marori and Enggano communities and speakers.

6.1 Working on Balinese as a native speaker: From apprenticeship to mentorship

My PhD research at the University of Sydney, later published as Arka (2003), marked the beginning of a four-year intensive co-creation process with non-Balinese linguists — particularly my supervisors, Professor William A. Foley and Professor Jane Simpson. This formative period was not just about completing a dissertation; it was a deeply collaborative academic apprenticeship that shaped me into an independent researcher. Beyond the dissertation, I co-authored several self-contained papers addressing key issues from the project. These included joint works with supervisors and other linguists, such as Wechsler and Arka (1998), Arka and Simpson (2008), and Arka and Manning (2008). Each collaboration brought new insights, methodologies, and disciplinary cross-fertilisation, pushing the boundaries of how Balinese grammar could be analysed and theorised.

Two critical lessons emerged during the early phase of my academic career. First, I became acutely aware of the limitations of native-speaker intuitions — a point also underscored by Ameka (2006). As native speakers, we often carry unconscious biases: we may privilege our own dialects, fail to account for variation, or overlook linguistic phenomena that seem self-evident and thus remain unarticulated. This experience taught me the importance of data representativeness, particularly the need to test linguistic judgments across speakers of different Balinese dialects. Additionally, working collaboratively with non-native linguists helped surface culturally significant distinctions and grammatical patterns that I, and other Balinese linguists, had previously under-analysed or overlooked.

This issue becomes particularly evident when examining traditional Balinese grammars written by fellow native-speaker linguists. For example, Oka Granoka et al. (1985) illustrates many of the limitations mentioned above, especially a tendency to overlook language-specific peculiarities, such as the existence of an undergoer voice distinct from the passive voice, which would likely be recognised if the linguist had a solid understanding of Balinese's typological profile within the Austronesian family. The treatment in Oka Granoka et al. is notably cursory — a feature common to many grammars produced under the auspices of Indonesian *Badan Bahasa*, which often follow a uniform, structurally top-down template rooted in basic structuralist linguistics. For example, chapters on syntax typically present basic constituent or grammatical structures — clauses, phrases, subjects, predicates, complements — accompanied by a single, simple example per structure. These examples appear to be self-elicited, lacking contextual grounding in actual language use, and notably, they are not drawn from any corpus of natural language use.

Although the grammar is written in Indonesian and aimed at both Balinese and non-Balinese Indonesian readers, the examples are presented without interlinear glosses, which hampers grammatical insight — only free translations are provided. Crucially, salient grammatical phenomena are often either under-described or completely absent. Consider the following contrasting examples, which the grammar uses to illustrate active (AV) and passive (PASS) voice:

(1) a. *Gegendong-e maek-in I Made Diarsa.*
beggar-DEF AV.close-APPL ART name
'The beggar approached I Made Diarsa.'
b. *I Made Diarsa paek-in-a baan gegendong-e.*
ART name close-APPL-PASS by beggar-DEF
'I Made Diarsa was approached by the beggar.'
(Oka Granoka et al. 1985: 86; glosses added)

While these examples do reflect a voice contrast, the grammatical analysis provided is superficial. The discussion does not address the fact that the suffix *-a*, glossed here as PASS, can also appear in structures that lack an overt PP. In such cases, *-a* is ambiguous, functioning as passive marker as in (1b) above, or as a bound pronominal clitic (=*a*), forming a morphosyntactically distinct construction known in Balinese as the Undergoer Voice or Objective Voice (UV/OV) (Arka 2003, 2008) — a voice type commonly found in Indonesian-type languages. Furthermore, the grammar omits any discussion of the *ka-* passive, which I (Arka 2003) have described as the "high passive."

Recent studies reveal that these *-a* and *ka-* passives coexist as integral components of the Balinese voice system, reflecting complementary functions with subtle distinctions shaped by semantics and social cognition (Arka & Mullan, to appear). For instance, the *-a* passive is generally preferred in contexts requiring agentivity or control (e.g., purposive clauses, negative imperatives), whereas the *ka-* passive, especially in its non-agentive or stative usage, aligns with less volitional, achieved-state readings (Udayana et al. 2025).

Another important lesson I recognised in my early academic career is the essential role of rigorous training across multiple strands of linguistics — including theoretical linguistics — for native-speaker linguists who wish to contribute meaningfully to the field. The ability to formulate precise research questions, engage critically with data, structure coherent arguments, and produce scholarly work — whether in presentations, publications, or theses — requires far more than intuitive knowledge. The earlier discussion of passive constructions in Balinese illustrates this point well: early grammatical descriptions by native-speaker linguists failed to capture the full complexity and typological significance of the language's passive constructions. In contrast, more recent work by fellow Balinese

linguists has begun to address these gaps — progress that has been enabled by advanced linguistic training. These developments clearly underscore the importance of such training for native-speaker linguists (see also Section 3.2).

Importantly, this kind of training involves more than theoretical instruction: it demands exposure to diverse linguistic models, academic writing conventions, and hands-on analytical skills, with computer-based technical proficiency. In today's digital context, this includes familiarity with widely used tools in language documentation and analysis, such as ELAN, FLEx, and Praat. In addition, for archiving purposes and following the "quartet" model proposed by Musgrave and Thieberger (2021), robust language documentation increasingly depends on interoperable, machine-readable formats that facilitate navigation and re-use across multiple documentation components. My own experience as a PhD student and as a team leader in ELDP-funded major documentation projects in Indonesia laid the foundation for this awareness and has since shaped how I mentor students and junior colleagues — particularly native speakers documenting their own languages, as discussed further in Section 5.

Now, as a senior scholar with over two decades of research experience on Balinese and other languages of Indonesia, I often engage in co-creation as a mentor, collaborator, and academic partner. Whether working with fellow Balinese linguists or non-Balinese colleagues, the co-creation process typically involves joint efforts in project conceptualisation, securing funding, developing methodologies, supervising data collection and analysis, and co-authoring publications and presentations. These collaborations are not hierarchical but reciprocal, drawing on our respective strengths and perspectives to produce work that is both theoretically robust and locally grounded. The lessons learned from working on Balinese continue to inform my approach in other research contexts, including my work with the Enggano community, which I turn to in the following subsection.

6.2 Working as an Indonesian Balinese linguist with Enggano and Marori speakers

In contrast to my experience as a native speaker working on Balinese, my engagement with Enggano (Austronesian) and Marori (Papuan) has placed me in the role of a non-native linguist, facing both familiar and unique challenges common to outsider researchers (cf. Ameka 2006). These challenges have been both linguistic and non-linguistic in nature, requiring thoughtful strategies, collaborative structures, and a flexible, ethnographically informed approach to ensure a meaningful and respectful documentation process.

Linguistically, the most immediate hurdle was the absence of native-speaker intuitions, which non-native linguists naturally lack. Unlike Balinese, where I

could draw on deeply internalised knowledge of structural variation and subtle semantic distinctions, in Enggano and Marori, I have had to learn the language from scratch — decoding forms, functions, and usage patterns through intensive collaboration with local speakers. A cornerstone of this process has been the involvement of dedicated local research assistants who then became CRCs, Mr. Engga Z. Sangian (for Enggano), and Mr Agus Mahuze and Mr Maximus Ndiken (for Marori). Their contributions have been essential. Over time, Engga in particular, has acquired basic linguistic training, enabling him to play a central role in interpreting grammatical nuances, identifying variation across dialects, and refining analytical distinctions. Together with other project team members, we have worked to develop a large spoken corpus, representing different villages and dialects of Enggano. This corpus — comprising naturalistic, transcribed, and translated texts — has become an indispensable resource for detecting structural patterns and meaning contrasts that neither outsider nor insider intuitions alone could fully capture.

One of the most rewarding aspects of this work has been the co-creative relationship between non-native and native researchers. Our team is deliberately structured to be complementary: my theoretical and methodological expertise is paired with local CRCs' deep cultural and linguistic knowledge. This collaboration helps mitigate the risks of both native-speaker blind spots (cf. Section 6.1) and outsider misinterpretation. For the Enggano project in particular, Engga's expanding skills — ranging from data recording and metadata management to transcription, translation, and even academic writing — have allowed him to co-author papers and contribute to publications that reflect both scholarly standards and community-based perspectives; Rajeg et al. (2024) and Sangian et al. (2024). His involvement exemplifies the transformative potential of training and capacity building, echoing the models discussed in Section 5.

For the Marori Project, the impact is promising and extends beyond linguistics. Long-term mentoring in my Marori documentation projects has enabled local CRCs, in particular Agus Mahuze, to transform technical skills in recording, transcription, and data management into broader leadership, project management, and community engagement capacities. These skills have sparked creative initiatives beyond linguistics, as seen in Agus Mahuze's establishment of *Mahuze Mandiri and Tunas Mandiri*, NGOs that combine cultural documentation with ecotourism and agricultural ventures in Kampung Wasur, securing funding, government support, and inspiring similar projects in the region (Arka 2025). For example, as seen in Figure 1, Kampung Wasur's local nursery showcases native flora, including wild orchids, reflecting the community's deep connection to the land and its traditions while drawing visitors from near and far. This trajectory — from skills transfer to innovation — illustrates how capacity building in documentation can generate lasting cultural and economic benefits for local communities.



Figure 1. Local garden featuring native orchids as a tourist attraction, with Agus pictured serving as a tour guide

My approach has also been ethnographic, recognising that grammatical forms and cultural practices are tightly interwoven. As Ameka (2006) argues, outsider linguists working in unfamiliar cultural settings are prone to missing culturally embedded meanings unless guided by community members. Local CRCs' participation has been crucial in bridging this gap: they not only explain linguistic usage but also situate it within local belief systems, rituals, and social norms. This kind of grounded interpretation cannot be gleaned from forms alone — it requires lived cultural experience. The project outcomes involving ethnographic insight include Hisa, Mahuze and Arka (2017; 2018), Arka et al. (to appear) and Rajeg et al. (2025), among others.

Importantly, my positionality as an Indonesian linguist of Balinese background has conferred certain advantages. Although I am not a native speaker of Enggano and Marori, my fluency in Indonesian, which functions as the lingua franca across the archipelago, facilitates smooth communication and collabo-

ration. Moreover, my familiarity with Indonesia's ethnolinguistic diversity has helped me build rapport and navigate local protocols with cultural sensitivity. Unlike foreign researchers, I do not face the same degree of cultural distance or bureaucratic restriction, allowing for sustained and reciprocal engagement with the community throughout the project lifecycle.[5]

In sum, the Enggano and Marori case studies demonstrate the value of co-creative, hybrid teams, where outsider and insider perspectives are not only combined but continually negotiated in the research process. The integration of linguistic training, corpus development, ethnographic insight, and mutual respect has enabled our projects to meet — and in many respects exceed — their original goals. This experience affirms the argument made throughout this chapter: that the most productive and ethically sound linguistic work emerges when native and non-native linguists collaborate as equal partners, drawing on their respective strengths to produce scholarship that is both academically rigorous and locally meaningful and useful.

7. Recommendations for practice, from extraction to co-creation: Strategies for ethical and effective collaboration

The foregoing sections clearly underline the point that to foster truly inclusive linguistic science, impactful language documentation research depends not just on scholarly rigour, but on cultivating co-creative, ethically grounded relationships with local communities. This means moving beyond short-term, outsider-driven extraction toward models of shared responsibility, local leadership, and strategic capacity building. In contexts like Indonesia — marked by linguistic, ethnic, and political complexity — these collaborative strategies must be finely attuned

5. A reviewer raised the point that being Indonesian can be a disadvantage in certain parts of Indonesia, particularly in Tanah Papua. I appreciate and fully acknowledge this important observation. In some areas — especially the central highlands — where Indonesian state presence is often perceived as colonial occupation, being Indonesian can indeed be a liability. In extreme cases, this has resulted in harassment and even lethal attacks on Indonesian civilians, including teachers, health workers, and, in other sectors, policemen and soldiers. While I have not personally experienced such hostility in Tanah Papua, nor am I aware of fellow Indonesian linguists encountering it, my knowledge is limited to colleagues — both Papuan and non-Papuan — who are typically affiliated with government institutions (e.g. universities, BRIN, or local branches of the National Language Board/Balai Bahasa). To my knowledge, none of these linguists have faced the kinds of threats or violence noted by the reviewer. Nonetheless, I recognise that such risks are real and must be taken into account when considering the positionality and safety of Indonesian researchers working in politically sensitive areas of Tanah Papua.

to local ecologies, which must also respect national language ideologies (Arka 2013: 74–76).

One foundational strategy is transparent expectation management from the outset. As noted in Sawaki and Arka (2018: 263–264), avoiding terms like *proyek*, which local communities often associate with state-driven aid programs, helps minimise misunderstandings. Instead, we use terms like *aktivitas* or *program* and host community meetings to explain project goals, timelines, and mutual benefits. These meetings are opportunities to build trust and include elders and younger, educated locals who may later become co-researchers or documentation leaders.

Crucially, recruiting and mentoring a capable local CRC has proven key to success. In my Marori work, Mr. Agus Mahuze became more than a consultant — he was a cultural liaison, logistics coordinator, and linguistic co-analyst (Arka 2018; Hisa et al. 2017, 2018). His long-term involvement ensured continuity and legitimacy within the community. As I argue in Arka (2025) and Arka and Mullan (to appear), empowering such local figures supports not only more nuanced and accurate analyses, but also community investment in the outcomes of the research.

Training, however, must be contextualised within the broader ecology of language endangerment, ethics and curriculum reform. As I note in Arka (2013: 81–82), minority language speakers often do not perceive language loss as an immediate threat. Instead, Indonesian is embraced as an "opportunity" — a tool for education, mobility, and survival. Language maintenance can seem counter-intuitive, especially to youth, unless it is explicitly linked to broader aspirations for wellbeing and identity continuity. In this light, capacity-building must address not only linguistic skills but also strategic motivation, helping communities understand how local language knowledge can coexist with, and even enhance, broader life goals. Furthermore, recent work highlights the importance of embedding decolonial and Indigenous research methodologies within linguistics curricula (Tsikewa 2021; Hudley, Montoya et al. 2024), and of reframing fieldwork as a collaborative, socially accountable practice, with institutional recognition of community-driven scholarship as central to the discipline. Building on these insights, I argue that incorporating community voices is essential at every stage — from project conception to implementation, including capacity-building training — to ensure sustainable local collaborations.

In line with this emphasis on sustainability and community collaboration, capacity-building must be paired with practices that ensure research is non-extractive and gives back to the community — particularly in relation to the materials collected. This includes clear communication from the outset about data ownership, long-term accessibility, and how materials will be used after the project's completion. In both the Marori and Enggano projects, local communities

were informed early on that all recordings and other data would be collected under strict ethical guidelines and archived in secure online repositories such as PARADISEC[6] and ELAR,[7] with the community retaining the right to decide levels of public access. In practice, only a few individuals expressed interest in receiving copies of recordings, often for personal reasons (e.g. as mementos of family members who had participated as language consultants). Nonetheless, both projects have made it a priority to leave copies of recordings and related data with local communities, where they are now managed by CRCs, reflecting our commitment to socially responsible, non-extractive research, in line with Hudley, Montoya et al.'s (2024) call for community members to act as long-term stewards and decision-makers in linguistic projects. In the Marori context, materials relating to ethnobotany and ritual sites have also proven valuable for non-linguistic purposes, particularly in supporting eco-tourism and Marori Cultural Centre initiatives led by Agus Mahuze. For Enggano, while we were unable to obtain the early 20th-century legacy recordings collected by German linguist Kähler (Kähler 1975), we were able to return copies of his legacy texts to the Enggano community.

A critical insight from Arka (2013) is that strategic language work in Indonesia is filtered through national ideologies like *Bhinneka Tunggal Ika* 'Unity in Diversity' and political constructs such as "threat" vs. "opportunity." Activities that appear benign or beneficial to linguists may be misinterpreted — especially in sensitive regions like Tanah Papua, where any assertion of cultural identity risks being classified as *ancaman* 'threat' by the state (Arka 2013: 85–86). While in principle political neutrality is desirable, in practice this is not always possible or even safe, particularly where research permits and long-term access depend on maintaining good relations with multiple stakeholders, including government actors. My own extended fieldwork experience in Merauke — a region with a strong government presence — has shown that early engagement with authorities (e.g. *Kesatuan Bangsa* (*kesbang*) 'National Unity' offices) and symbolic gestures — like displaying presidential photos in workshops — can be crucial for securing institutional support and minimising political friction. At the same time, I recognise that in other areas, particularly in parts of the central Papuan highlands where armed conflict persists, such displays could have the opposite effect and place researchers at risk. In these contexts, the safest and most productive strategy is to follow local advice on the best approach for maintaining long-term engagement not only with the speech community but also with local authorities, with the aim of gaining their support for work that ultimately benefits the community.

6. https://www.paradisec.org.au/

7. https://www.elararchive.org/

Finally, the choice of theoretical framework also plays a crucial role. As noted earlier (cf. Arka & Mullan, to appear), modular architectures such as LFG offer the flexibility to incorporate culturally grounded concepts — such as Balinese speech levels and deference or politeness markers — within a formal analytical model. However, one of the strengths of LFG is that, due to its modular design, it can be adopted in a simplified form, allowing linguists to capture rich, language-specific insights without explicating or relying on the full formal apparatus. This adaptability makes it especially valuable in collaborative settings: it enables native-speaker collaborators — even those without formal linguistic training — to recognise their intuitions and cultural knowledge in the resulting grammatical description. In turn, this strengthens their sense of ownership and fosters deeper engagement with the research process.

In sum, successful collaboration in language documentation — especially for non-local linguists — requires a dynamic interplay of cultural sensitivity, institutional negotiation, theory-practice integration, and capacity building. These strategies help transform documentation research from a one-time research intervention into a sustainable, community-driven process of knowledge preservation and revitalisation.

8. Conclusion

This chapter has explored how collaborations between native- and non-native-speaker linguists can diversify and deepen linguistic science. Drawing on personal experience and institutional reflections from over two decades of research in Indonesia and echoing ideas first by Ameka (2006) and also later by others (Florey & Himmelmann 2010:122; Dobrin & Berson 2011; Czaykowska-Higgins 2009; Bischoff & Jany 2018, among others), I have argued that inclusive, co-creative models of linguistic description are not only possible but essential — particularly in linguistically rich, socially complex, and historically under-resourced contexts like Indonesia.

Several key lessons have emerged. First, the success of linguistic documentation efforts depends fundamentally on rethinking who is positioned as a knowledge producer. Native speakers are not merely data sources; when equipped with the right tools, training, and support, they become co-analysts, co-authors, and stewards of their linguistic heritage. Their insights not only enhance the descriptive depth of linguistic work but also ground it in culturally embedded, socially meaningful knowledge systems.

Second, rigorous linguistic training — whether field-based or academic — plays a transformative role in enabling local collaborators to move from partic-

ipation to co-creation and leadership. This training includes not just linguistic instruction (from descriptive to typological and theoretical) but also hands-on experience with analytic tools, scholarly writing, and interdisciplinary perspectives. As demonstrated through the Balinese, Marori, and Enggano case studies, such capacity-building efforts yield outcomes that are richer in both scientific and community impact.

Third, the role of non-native linguists remains significant — but not as extractive outsiders. Rather, they function best as facilitators, collaborators, and bridge-builders — particularly where local capacity is still emerging. When practiced ethically and reflexively, non-native participation can help catalyse long-term, locally anchored research ecologies. The experiences shared here also affirm that field-based training and mentorship models — especially those fostering continuity through local research assistants/collaborators or graduate students — are vital for embedding documentation within communities over time.

Finally, while acknowledging that research is inherently non-neutral (cf. Czaykowska-Higgins 2009), this chapter has underscored that linguistic research today must be reimagined as a co-constructed, socially accountable practice. This view, increasingly shared across the discipline (cf. Tsikewa 2021, Hudley, Mallinson & Buchholtz 2024), aligns with broader shifts in modern documentary and field-based linguistics: away from extractive paradigms and toward sustained, reciprocal partnerships that reflect a plurality of epistemologies. Such a shift requires structural changes — not only in how projects are designed and funded, but also in how institutions recognise and support the long-term engagement needed to produce collaborative, ethical, and impactful research. This movement toward decolonizing research praxis ensures that community voices are central to linguistic knowledge production, echoing Hudley, Mallinson and Buchholtz' (2024) call for fundamental transformations to the structures, incentives, and epistemologies that shape the discipline.

In short, the lessons learned from the Indonesian context offer a compelling model for the emerging paradigm in linguistic science: one that is pluralistic, interdisciplinary, co-creative, and grounded in mutual respect. As we move forward, diversifying science must mean more than increasing representation; it must entail transforming the very practices through which knowledge is made — ensuring that the voices and values of speaker communities are central to how we document, analyse, and understand the world's languages.

Acknowledgements

Dedicated to Nikolaus Himmelmann, whose pioneering work in language documentation and capacity building in Indonesia — and in Austronesian linguistics more broadly — has been a lasting inspiration. I have been fortunate to work with him, including in the UBUD Documentation Workshops (2006–2007) and during my Humboldt Fellowship at the University of Cologne, which he kindly sponsored. I also thank the anonymous reviewers, Sonja Riesberg (editor), Charbel El-Khaissi, and Amrita (Mita) Nugraheni for their helpful comments on this chapter.

References

Ameka, Felix K. 2006. "Real Descriptions: Reflections on Native Speaker and Non-native Speaker Descriptions of a Language." In *Catching Language: The Standing Challenge of Grammar Writing*, ed. by Felix K. Ameka, Alan Dench, and Nicholas Evans, 69–112. Berlin: Walter de Gruyter.

Arka, I Wayan. 2003. *Balinese Morphosyntax: A Lexical-functional Approach*. Canberra: Pacific Linguistics.

Arka, I Wayan. 2007. "Local Autonomy, Local Capacity Building and Support for Minority Languages: Field Experiences from Indonesia." In *Documenting and Revitalizing Austronesian Languages*, ed. by D. Victoria Rau, and Margaret Florey, 66–92. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/1353

Arka, I Wayan. 2008. "Voice and the Syntax of =a/-a Verbs in Balinese." In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter K. Austin, and Simon Musgrave, 70–89. Stanford: CSLI.

Arka, I Wayan. 2010. "Ritual Dance and Song in Language Documentation: Vera in Rongga and the Struggle over Culture and Tradition in Modern Manggarai-Indonesia." In *Language Endangerment in the Austronesian World: Challenges and Responses*, ed. by M. Florey, 90–109. Oxford: Oxford University Press.

Arka, I Wayan. 2013. "Language Management and Minority Language Maintenance in (Eastern) Indonesia: Strategic Issues." *Language Documentation & Conservation* 7: 74–105. http://hdl.handle.net/10125/4568

Arka, I Wayan. 2015. "On the Dynamics of Glocalisation and Minority Language Conservation in Contemporary Indonesia." In *International Conference on Language, Culture and Society*, ed. by Obing Katubi, 1–10. Jakarta: LIPI.

Arka, I Wayan. 2018. "Reflections on the Diversity of Participation in Language Documentation." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 132–139. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24815

Arka, I Wayan. 2025. "Language Documentation and the Multi-Dimensionality of Capacity Building: Framing Research Diversity in an Indonesian Ethno-ecological Context." In *Projecting Voices: Essays in Language and Linguistics in Honour of Jane Simpson*, ed. by C. O'Shannessy, J. Gray, and D. Angelo, 1037–1061. Canberra: Asia-Pacific Linguistics.

Arka, I Wayan, and Keira Mullan. To Appear. "Social Cognition, Typology, and Modularity in Linguistic Theory: An Indonesian Perspective." In *Linguistic Worlds and Social Minds: What Language Documentation and Linguistic Diversity Can Tell Us about the Nature and Origin of Language*, ed. by I Wayan Arka, Barth Danielle, and Henrik Bergqvist.

Arka, I Wayan, Arono, Dendi Wijaya, and Engga Zakaria Sangian. To appear. "The Socio-ethnolinguistic Landscape of Contemporary Enggano." In *Enggano: Historical and Contemporary Perspectives*, ed. by I Wayan Arka, Mary Dalrymple, and Charlotte Hemmings. London: EL Publishing.

Arka, I Wayan, and Christopher Manning. 2008. "Voice and Grammatical Relations in Indonesian: A New Perspective." In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter K. Austin, and Simon Musgrave, 45–69. Stanford: CSLI Publication.

Arka, I Wayan, and Jane Simpson. 2008. "Objective Voice and Control into Subject Clauses in Balinese." In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter K. Austin, and Simon Musgrave, 90–127. Stanford: CSLI Publication.

Bischoff, Shannon T., and Carmen Jany (eds.) 2018. *Insights from Practices in Community-Based Research, From Theory To Practice Around The Globe*. Berlin, Boston: De Gruyter Mouton.

Briggs, Charles L. 1986. *Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research*. Cambridge: Cambridge University Press.

Cargill, Margaret, and Sally Burgess. 2017. "Introduction: Unpacking English for Research Publication Purposes [ERPP] and the intersecting roles of those who research, teach and edit it" In *Publishing Research in English as an Additional Language: Practices, Pathways and Potentials*, ed. by Margaret Cargill, and Sally Burgess, 1–11. Adelaide: University of Adelaide Press.

Czaykowska-Higgins, Ewa. 2009. "Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities" *Language Documentation & Conservation* 3: 15–50. http://hdl.handle.net/10125/4423

Dobrin, Lise M., and Josh Berson. 2011. "Speakers and Language Documentation." In *The Cambridge Handbook of Endangered Languages*, ed. by Peter K. Austin, and Julia Sallabank, 187–211. Cambridge: Cambridge University Press.

Dobrin, Lise M., and Saul Schwartz. 2016. "Collaboration or Participant Observation? Rethinking Models of 'Linguistic Social Work.'" *Language Documentation & Conservation* 10: 253–257.

Enfield, N. J. 2002. *Ethnosyntax: Explorations in Grammar and Culture*. Oxford: Oxford University Press.

Evans, Nicholas, and Alan Dench. 2006. "Introduction: Catching language." In *Catching language: The Standing Challenge of Grammar Writing*, ed. by Felix K. Ameka, Alan Dench, and Nicholas Evans, 1–39. Berlin: Walter de Gruyter.

Florey, Margaret, and Nikolaus Himmelmann. 2010. "New Directions in Field Linguistics: Training Strategies for Language Documentation in Indonesia." In *Endangered Languages of Austronesia*, ed. by Margaret Florey, 121–140. Oxford: Oxford Univesity Press.

Himmelmann, Nikolaus. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Hisa, La, Agustinus Mahuze, and I Wayan Arka. 2017. "The Ethnobotanical-linguistic Documentation of Sago: a Preliminary Report from Merauke." *Linguistik Indonesia* 35 (2): 187–200.

Hisa, La, Agus Mahuze, and I Wayan Arka. 2018. *Etnobotani: Pengetahuan lokal suku Marori di Taman Nasional Wasur Merauke*. Merauke: Balai Taman Nasional Wasur Merauke.

Hudley, Anne H. Charity, Christine Mallinson, and Mary Bucholtz (eds). 2024. *Decolonizing Linguistics*. New York: Oxford University Press.

Hudley, Anne H. Charity, Ignacio L. Montoya, Christine Mallinson, and Mary Bucholtz. 2024. "Introduction: Decolonizing Linguistics." In *Decolonizing Linguistics*, ed. by Anne H. Charity Hudley, Christine Mallinson, and Mary Bucholtz, 1–45. New York: Oxford University Press.

Kähler, Hans. 1975. *Texte von der Insel Enggano. (Berichte über eine untergehende Kultur)*. Berlin: Dietrich Reimer.

Kroeger, Paul. 2007. "LFG as a Framework for Descriptive Grammar." In *Proceedings of LFG07*, ed. by M. Butt, and T. H. King. Stanford: CSLI.

Montrul, Silvina. 2016. *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press.

Musgrave, Simon, and Nick Thieberger. 2021. "The Language Documentation Quartet." *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages* Volume 1 (Papers): 6–12. Stroudsburg, PA: Association for Computational Linguistics.

Oka Granoka, Ida Wayan, Ida Bagus Udara Naryana, I Wayan Jendra, I Nyoman Medra, I. G. N. Putra Yasa, and I Wayan Bawa. 1985. *Tata Bahasa Bali*. Denpasar: Proyek Pengembangan Bahasa dan Sastra Indonesia dan Daerah Departemen Pendidikan dan Kebudayaan.

Polinsky, Maria, and Olga Kagan. 2007. "Heritage Languages: In the 'Wild' and in the Classroom." *Language and Linguistics Compass* 1 (5): 368–395.

Rajeg, Gede Primahadi Wijaya, Charlotte Hemmings, I Wayan Arka, and Engga Zakaria Sangian. 2024. "Enggano Middle Voice and Valence Over Time." Presentation, 15th International Conference of the Association for Linguistic Typology (ALT XV), Nanyang Technological University, Singapore, 4–6 December, 2024.

Rajeg, Gede Primahadi Wijaya, Charlotte Hemmings, Engga Zakaria Sangian, Dendi Wijaya, and I Wayan Arka (with Milson Kaitora, Harun Kaharubi, M. Raflizen Kaitora (alm.), Aron Kaitora (alm.), Johansen Kaharubi, Ishar Timius Kaitora, Marlansius Kaharubi, Adam Kurniawan Kauno, & Resiawati Kaitora). 2025. *Kamus Bahasa Enggano* (*1st ed*). Oxford: University of Oxford & Zara Abadi.

Sangian, Engga Zakaria, Charlotte Hemmings, Arono Arono, Gede Primahadi Wijaya Rajeg, and I Wayan Arka. 2024. *Pihia' Yic Enggano VII*. Oxford: University of Oxford & Zara Abadi.

Sawaki, Yusuf, and I Wayan Arka. 2018. "Reflections on Linguistic Fieldwork and Language Documentation in Eastern Indonesia." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 256–266. Honolulu: University od Hawai'i Press. http://hdl.handle.net/10125/24827

Tsikewa, Adrienne. 2021. "Reimagining the Current Praxis of Field Linguistics Training: Decolonial Considerations." *Language* 97 (4): e293–e319.

Udayana, I Nyoman, G. P. W. Rajeg, and Ida A Made Puspani. 2025. "The (Non)canonical Status of the ka- Passive in Balinese." *Linguistics Vanguard*: 1–15.

Wechsler, Stephen, and I Wayan Arka. 1998. "Syntactic Ergativity in Balinese: An Argument Structure Based Theory." *Natural Language and Linguistic Theory* 16. 387–441. https://www.jstor.org/stable/4047955.

Woodbury, Anthony C. 2003. "Defining Documentary Linguistics." In *Language Documentation and Description*, ed. by Peter K. Austin, 35–51. London: SOAS.

Woodbury, Anthony C. 2011. "Language Documentation." In *The Cambridge Handbook of Endangered Languages*, ed. by Peter K. Austin, and Julia Sallabank, 159–185. Cambridge: Cambridge University Press.

CHAPTER 5

From the audible to the meaningful

The role of listener/speakers in making sense of 60 years of Dalabon recordings

Nicholas Evans & Manuel Pamkal
Australian National University | Dalabon Aboriginal Cultural Experiences

The documentarist turn remains lopsided with respect to the form<>meaning mapping: we risk knowing what speakers said, but not what they meant.

Original recordings thus need to be supplemented by discussions with speakers that use methods more similar to traditional methods of textual hermeneutics.

Here we exemplify these problems, drawing on our own work with previously untranscribed, untranslated audio recordings in Dalabon. We examine four key issues: an unreported "reversal" construction; adding senses to known vocabulary; unearthing previously unrecorded new food-preparation vocabulary; reference of "triangular kin terms" in actual conversation.

Such interpretive discussions with later-generation community members reach beyond the audible to increase the reach of semantic interpretation.

Keywords: Dalabon, hermeneutics, language documentation, reversal constructions, triangular kin terms, field semantics

> The net should be cast as widely as possible. That is, a language documentation should strive to include as many and as varied records as practically feasible … Ideally … a language documentation would cover all registers and varieties, social or local; it would contain evidence for language as a social practice as well as a cognitive faculty; it would include specimens of spoken and written language; and so on.
>
> (Himmelmann 2006: 2)

https://doi.org/10.1075/slcs.240.05eva
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

1. Introduction

Nikolaus Himmelmann has made many contributions to the field, from Austronesian to Papuan linguistics, from typologies of word class and voice to prosody and much more besides. But it is arguably his theorisation of the new subfield of language documentation, through such seminal publications as Himmelmann (1998, 2006) and Gippert et al. (2006), that has had the most far-reaching impact on how the field of descriptive and typological linguistics is practiced today. His clear-sighted setting out of how we should go about documenting language, in a way that maximises the utility of our materials, for future generations who may ask completely different questions about the language to those we ask from behind the inevitable myopia of our current world-view, has driven massive advances in the global effort to record, document and archive our fragile linguistic heritage while this is still possible. This in turn has been fundamental in leading to the establishment of such important programs as DobeS (*Dokumentation bedrohter Sprachen*) funded through the Volkswagenstiftung, and the threefold program of ELDP (*Endangered Languages Documentation Program*), ELAR (*Endangered Languages Archive*) and ELAP (*Endangered Languages Academic Program*, now defunct) funded through Arcadia, administered first by SOAS and more recently by the Berlin-Brandenburg Academy of Sciences and Humanities. As a result of this, we now have many thousands of hours of archived material from many hundreds of languages all over the world.

However, there is an achilles heel in the documentarist program. Constraints of time, workflow and funding-agency priorities mean that corpora will almost always have a "Russian doll" structure: the number of recorded hours far exceeds the number of hours transcribed, which in turn far exceeds the number of hours for which we have careful, sensitive semantic treatments.

The first discrepancy reflects the "transcription bottleneck" which has traditionally meant ratios of between 30:1 and 100:1 of transcription time to recording time (Durantin et al. 2017; Foley et al. 2018). This is because of the difficulties of hearing and transcribing a language which the investigator is not yet fluent in, especially in situations where there is nobody from the speech community able to carry out transcription themselves, e.g. when all remaining speakers are non-literate or from a generation not used to keyboard technologies. And the more naturalistic the material — which is precisely the goal which the documentarist program has been encouraging us to aspire to — the more "messy" the transcription task we are faced with. (Transcribing elicitation can go quite fast, whereas transcribing multi-party and perhaps even multilingual conversations with many overlaps can take even longer than the 100:1 ratio mentioned above.)

The second discrepancy reflects the "asymmetry of sound and meaning in the documentation process", to quote a formulation flowing from another current of the Cologne linguistic tradition (Evans & Sasse 2007: 67). Typically, transcriptions will be furnished with first-pass rough translations, and interlinear glosses, but these are far from answering the question "what did the speaker really mean?" There will often be words that are encountered for the first time in the documentary engagement with the language, and for which translations might range from 'type of plant', where there is partial comprehension of their meaning, to '??', where there is none at all. Or there will be a number of "false positives", from a semantic point of view, where the investigator has encountered the word or morpheme before, thinks they know what it means, and translates and/or glosses accordingly. And even where a translation does make sense at some level, e.g. 'he arrived' or 'over there', this is not the same as achieving a full referential understanding, e.g. 'Nikolaus Himmelmann arrived' or 'in the north-east corner of the Meister-Ekkehart-Straße Institute main linguistics library room'.

So if we are to fully understand that most basic function of language, namely to convey meaning, our work is incomplete until our translations have been lifted to this level of precision. It is this fine level of transcription and translation that we should be aiming for in language documentation.

Thus, whatever program of documentation workflow we embark upon for a given language, it is vital that we put an adequate portion of effort into detailed discussions of meaning. In this article we will show how, even in quite advanced stages of language loss where just one or two fluent hearer/speakers remain, it is possible to revisit much earlier recordings and obtain a vastly enriched body of materials with which to make sense of the language.

Dalabon, the subject of this article, is one of around ten languages of the Gunwinyguan language family (Evans & Marley 2023), centred on the escarpment country of Western and Central Arnhem Land in Northern Australia. There has been a substantial amount of published work on Dalabon, including numerous articles, two PhD theses, two sets of deposited corpus materials in ELAR, and a preliminary dictionary (see Evans 2021 for a survey of publications). Unfortunately, this language, described as still vigorous in the nineteen sixties by Capell (1962), has now dwindled to the point where there are just a few fluent speakers; Manuel Pamkal is arguably the most fluent.[1]

1. You can get an idea of what Dalabon sounds like from two online resources: the Gerlingo Dalabon site (https://www.gerlingo.com/language_detail.php?langID=7) contains nine glossed and translated texts with streamed audio files, and the Dalabon Language Youtube Channel *Ngarrahyangwurlkiyan* (https://www.youtube.com/channel/UCDUFxXg2mG564H9g8BBr7BQ) contains six Dalabon videos with subtitles in Dalabon and English.

For Bininj Kunwok (Evans 2003), a closely-related language, recent work by Alexandra Marley has been able to construct a corpus spanning a century (by speaker birthdate) through the careful integration of early and more recent recordings — see Marley (2020). This proved enormously useful in understanding fine-grained changes in the language through time. For Dalabon several linguists have been engaged for some years in working our way through recorded materials, going back to the 1960s: in addition to the two present authors, Salome Harris and Sarah Cutfield have also done substantial work on this task. From among these materials, we focus in this article on the five recordings summarised in Table 1.

Table 1. Recordings drawn on in the present text

Speaker(s)	Year	Recorder	Archival name (media file)	Transcription notes
Jack Djaŋaḏum (= Chadum/ Chatham), Tommy Rindjir	1962	LaMont West	West61-Q408A4 4 mins 07secs	Transcribed by NE and MP, 7–8/4/2018, Hot Springs, Katherine
Copper Marnbudji	1962	LaMont West	WEST61-Q401A4 10 mins 23 secs	Transcribed by Salome Harris, Sarah Cutfield and NE between 2010 and 2024
Nabardayal Nadjimerek	1963	Joy Kinslow-Harris*	KINSLOW-HARRIS_J03-000320A.TMP 4 mins 03 secs	Transcribed by NE and MP, Katherine, 15–16/ 3/2022
Bob Barinjbungu	1969	Joy Kinslow-Harris**	Dalabon1969BobBariynbunguBoosted 28 mins 24 secs	Transcribed by NE with Manuel Pamkal in Katherine from 15/2/ 2019–3/6/2024.
Winnie Baruwei, David Jentian Nangan.golod, Jack Chadum and others	1974	Barry Alpher	ALPHER_B02-003393A.wav (AIATSIS Archive) 14 mins 53 secs	Initial transcription by Salome Harris (2009–2010), with full checking by NE & MP in 2023.

* Fuller AIATSIS archival details: Rec. Joy Kinslow-Harris, July-Aug 1965, Oenpelli and Goulburn Island. Woman is Lena (no further details given) and man is Lofty. Collection Kinslow-Harris_J03, Tape/Track 00320a.

** Fuller AIATSIS archival details: Archive Tape A 174 Track A, Field Tape Number 2, Side 1.

The rest of this article is structured as follows. In Section 2 we focus on a particular point of grammatical interpretation, where recordings from three different speakers in Table 1 all use an unanticipated "reversal" construction in which the

direction of action between subject and indirect object is the opposite of what all previous work has led us to expect. In Section 3 we tackle the question of how careful translation can add semantic nuances to terms which have already been recorded. In Section 4 we use the semantic domain of cooking terms to show how substantial amounts of new vocabulary can be unearthed through careful listening to old recordings, by speakers who may no longer use these terms actively but can still understand and explain their meaning. In Section 5 we show how it is possible, in cases where listeners combine sufficient background knowledge of the people involved with a delicate use of elaborate kinship terminology, to go beyond knowing what it means, to working out who it is referring to in the interpretation of a particular text — something we need to be able to do if we are to fully understand how a language is used in real conversation. We conclude in Section 6 with some brief programmatic remarks.

2. Grammar and interpretation

In documenting a language and preparing our analyses, we sometimes reach a "plateau stage" where we have inventoried a set of morphemes and believe we understand their functions. In the case of the pronominal prefixes for subject and object/indirect object, our[2] analysis had already stabilised by Evans et al. (2001), which sets out all logical combinations of subject and object. Subsequent work (Evans 2013) focussed on some of the formal complexities that arise when the "basic" subject+object combinations interact with a variety of tense, aspect and mood categories to spawn a number of extra paradigms.

The danger of this plateau stage is that we can fail to detect exceptions where the standard analysis doesn't hold. This is particularly likely to happen if we proceed to analyse recorded, transcribed but untranslated texts by applying the principles which we have worked out in previous research. It is at this point that getting careful translations by native speakers prove vital in showing where our existing models are inadequate. I will illustrate this, for the recordings above, by considering the problematic problem of "reversals" — where a pronominal prefix combination that we would expect to mean 'X acting upon Y' in fact means the reverse, 'Y acting upon X'. I take the term from the recent survey by Baerman (2007: 33), who defines it (and situates it within the field of other terms sometimes used):

2. Here I am using "our" in the sense of the collective scholarly consensus, which included published work by A. Capell (1962) and Barry Alpher (1982), as well as fieldnotes by myself and Francesca Merlan, and such articles as Evans et al. (2001).

> morphological reversal describes the situation where the members of a morphological opposition switch their functions in some context … There is a long tradition of polemic against the notion that morphology can encode systematic reversals, and an equally long tradition of reintroducing them under different names (e.g. polarity, exchange rules or morphosyntactic toggles).

First some background: Dalabon has two sets of prefixes, monovalent (subject only) and divalent (subject and object/indirect object). A sample of prefixes (sometimes involving proclitic + prefix combinations) is given in Table 2. Given that there are 102 distinct subject/object combinations (Evans et al. 2001), and that each of these cells is potentially multiplied by five different TAM-categories (realis, irrealis, hortative, apprehensive, and purposive) it would take us too far afield to show them all here — see Evans et al. (2001) and Evans (2013) for details. But the forms shown here will suffice to understand the issues in this paper. Note that for some combinations, the second argument is realised by a proclitic, through there is a great deal of variation in whether this is realised as a separate word or reduced and fused with the subject prefix, e.g. 2SG.A>3PL.(I)O can be either *bulu da-* or fused to *bulda-* (see Evans et al. 2008 for further examples).

Table 2. Selected pronominal prefixes

Subject person/ number	Monovalent	Divalent with 3SG.(I)O	More complex divalent combination	Divalent form
1SG	*nga-*	*nga-*	1SG.A>3PL.(I)O	*bulu nga-*, *bulnga-*
1DU.DIS	*nge-*	*nge-*	1DU.DIS.A>3PL.(I)O	*bulu nge-*, *bulnge-*
1*PL	*yala-*	*yila-*, *yala-*		
12PL	*ngarra-*	*ngûrra-*, *ngarra-*		
3SG	*ka-*	*ka-*, *bûka-***	3SG.A>12DU.(I)O	*njeh ka-*
			3SG.A>1DU.(I)O	*njerr ka-*
			3SG.A>12PL.(I)O	*ngorr ka-*
2SG	*dja-*	*da-*	2SG.A>3PL.(I)O	*bulu da-*, *bulda-*

* In non-singulars, 1DU/PL indicates exclusive, and 12DU/PL indicates inclusive.

** *Ka-* is used whenever the object is inanimate, or lower than the subject (e.g. a human acting upon a dog). *Bûka-* is used whenever the object is human, or when an indirect object rather than a direct object is involved. The variant *biya-* is sometimes found instead of *bûka-*.

In general, monosyllabic monovalent prefixes can also be interpreted as divalent prefixes where the object is third person (thus *nga-* '1SG.S; 1SG.A>3SG.(I)O'. For disyllables, some speakers shift the first vowel from *a* to *i* (after palatals) or *u/û* (elsewhere) in forming the divalent third-person singular object form from the monovalent form: cf. *yala-* '1PL.S' but *yila-* '1PL.A>3SG.(I)O'. Other speakers do not make the vowel change in forming the divalent series. Where the object/indirect object is other than 3rd person singular, things get more complex: In some cases, a special clitic pronoun is used (e.g. *njerr* '1DU', as in *njerr kah-nang* 'he saw the two of us') while in others (particularly involving the first person singular) third-person singular object forms are recycled.

Armed with this information, we can approach some actual examples of standard uses, taken from Evans et al. (2004). We will use examples where the indirect object is introduced by the benefactive applicative *marnû-*, since as we shall see this is the construction type in which, so far, all the reversed examples of interest to us are found.

First consider what happens with first person singular indirect objects. The form *ka-* has a number of meanings in Dalabon: its basic monovalent use is as third person subject, but in divalent constructions it can be used for third person subject acting on third person inanimate object, third person subject acting on first person object, and second person subject acting on first person object. Example (1) illustrates third person singular acting upon first singular.

(1) *Nga-h-munkuwa-nj, ko kanh mak ka-marnû-dudj-mi.*
1SG.A>3SG.O-R-send-PP so.that DEM NEG 3SG.A>1SG.(I)O-BEN-return-IRR
'I've kicked him out, so he won't come back to me' (Evans et al. 2004: 224–225)

Now consider what happens with first person non-singular (derived indirect) objects. For these an appropriate object clitic is used. An example is:

(2) *Mak ngorr ka-marnu-dudjm-iyan.*
NEG 12PL.(I)O 3SG.A-BEN-return-FUT
'He'll never come back to us.' (Evans et al. 2004: 226)

Where the argument introduced by the benefactive is third person singular, the same forms are used as with third person singular objects of regular transitive verbs, following the formation rules outlined above (and in more detail in Evans et al. 2001). Two examples are (3), with the first person singular subject *nga-*, and (4), with the first person exclusive plural subject *yila-*.

(3) *yeah, yawok nga-h-marnû-djird-djirdma-nginj*
yeah cheeky.yam 1SG.A>3SG.(I)O-R-BEN-ITER-steal-PP
'Yeah, I stole lots of cheeky yams from him.'
[Dal9605DaisyBorduk 00:03:42–45]

(4) *Kanunh modjdjanjno yila-h-marnû-wa-n.*
DEM custom 1PL.A>3SG.(I)O-R-BEN-follow-PRS
'We follow that custom with regard to it (the long-horned grasshopper).'
(Evans et al. 2004: 228)

So far, all these involve "standard" uses of the pronominal prefixes. These standard uses characterise all speakers I have worked with in person, as well as those who appear in materials on Dalabon published by other investigators (Cutfield 2011; Ponsonnet 2013, 2014).

However, in going over the older recordings discussed here, in each one of them we find examples of what we might call the "reversed benefactive construction". In this, the mapping of subject and indirect object person/number values onto the respective pronominal prefixes is reversed compared to what we expect. Starting with (5), *ngah-marnu-* means 'it [will come] for me' instead of the expected 'I [will come] for it', and *yila-marnu-* in (6) means 'it [puts itself there] for us' instead of the expected 'we … for it' (similarly in (7)), *ngah-marnu-* in (8) means 'he [died] on me' instead of expected 'I [died] on him/her', and *ngarrah-marnu-* in (9) means 'it [is there] for us' instead of the expected 'we … for it'.[3]

(5) [Preceding lines, translated (11:47–12:05): 'We're all out of tobacco, we're starving for tobacco; I'll walk, I've got no money; nothing. But it would be good if they gave us some of the fine tobacco to pack in. OK, I've got nothing in hand, I'm just a beggarman poor feller me.']
but ngan worrayh, bad ke buhkorre-kuno kardû nekwik
but 1SG.POSS alright but PART soon-TIME maybe next.week
nga-h-marnû-burlhm-iyan
1SG.A>3SG.(I)O-R-BEN-arrive-FUT
↜3SG.A>1SG.(I)O
'But I'm alright, (pension) money, you know, should be coming up sometime soon, maybe next week it'll come to me.'
[Ngalk1974. Winnie Baruwei 12:06–12:13]
[Following lines, translated (12:14–12:22): 'It was that whitefella who told me, whatsisname, Jerry, it was Jerry who told me, tobacco, yeah, tobacco.']

3. For this and other glosses in this section, we give the standard glosses in terms of directionality on the first line, with a "reversed" gloss directly underneath, preceded by the symbol ↜; each special reversed reading is also discussed in the main text. Thus we gloss this example on the first line as 1SG.A>3SG.(I)O 'I acting upon him/her', then directly underneath it as ↜3SG.A>1SG.(I)O.

(6) [Preceding lines, translated (2:26–2:45): 'The two of us, with our dad, our dreaming, as water,']
drimin *yila-marnu-yu-rr-inj*
dreaming 1PL.A>3SG.(I)O-BEN-put-RR-PP
ꟻ3SG.A>1PL.(I)O
'The dreaming put itself there for us.'
[Copper_Marnbudji_Merged_NE2016[4] 2.46–2.51]
[Following lines, translated (2:51–56): 'We two called it *dodoyh* (great-grandmother/auntie), we called her *djongok* (auntie).']

(7) [Follows on immediately from (6) (so 2:57–3:01): 'Our father made himself into a rock,']
wah *yila-h-marnu-kolh-yu-rr-inj*
water 1PL.A>3SG.(I)O-R-BEN-water-put-RR-PP
ꟻ3SG.A>1PL.(I)O
'He/it put itself there (in the ceremony) as water for us.'
[Copper_Marnbudji_Merged_NE2016 3:01–3:04]
[Following lines, translated (3:04–3:09): 'NgalBadbulu made the ceremonial bones stand up at that place.']

(8) [Preceding lines, translated (3:40–3:45): 'My father, when he died…']
bulu-ngan *nga-h-marnû-donj* *na-Mandolmû*
father-1POSS 1SG.A>3SG.(I)O-R-BEN-die-PP M-[personal.name]
ꟻ3SG.A>1SG.(I)O
'My father died on me, na-Mandol.'
[Copper_Marnbudji_Merged_NE201 3:45–3:47]
[Following lines, translated (3:47–3:51): 'Yes NaMandol died, my father.']

(9) [Preceding lines, translated (3:24–3:51): 'From up in the escarpment country, coming downstream, from Malkakah, from Manbal, from Marranguy, from Ngurrbburl, from Yikandjalalk, from Kaddenj, from up in the high country, at Maluka, leaving the cycad palms, a water place, just now that we had been getting really thirsty:']
ngaleh *ngarra-h-marnû-kolh-yu-ø*
hey 12>3SG.(I)O-R-BEN-water-lie-PRS
ꟻ3SG.A>12PL.(I)O
'"Oh, there's water here for us!"' [Bob Barinjbungu 00:03:52–55]
[Following lines, translated (3:52–4:01): 'We were all saying, there from the high country, from the rock country, through a narrow place, from the gorge there.']

4. The final form of the transcription of this text reflects successive transcriptions by Salome Harris, Sarah Cutfield and Nicholas Evans, working with Manuel Pamkal.

It is not clear at this stage how these exceptions are to be accounted for. Do they represent an older stage of the language, in terms of how the pronominal prefixes work? It is worth pointing out that in related languages, such as Bininj Kunwok, the reversal of action flow in such cases would usually be shown by adding a suffix *-n* to the base form of the prefix, thus (in Bininj Kunwok) *nga-marne-yolyolmeng* 'I said to him' but *ngan-marne-yolyolmeng* 'he said to me'. Could it be that a system like this, widespread in Gunwinyguan languages, got neutralised in an earlier phase of Dalabon through the loss of *-n* in this environment, creating the ambiguity found in the speakers exemplified above, who formed the last gasp of an earlier system, and that this was then refashioned into the new (and logically rather different) system described in analyses based on younger speakers, such as that presented in Evans et al. (2001)? Or could there be some subtle additional factors, such as discourse factors allowing this benefactive reversal?

These questions remain unanswered. But the point of relevance to the present article is that it was only possible to detect these discrepancies thanks to the input of careful listening by an actual speaker, who was able to offer a translation that runs in the opposite direction to what we would expect. The fact that the same construction, always with a benefactive, is found across three speakers in this "older" corpus, clearly indicates that we are not dealing with speech errors but with a clear pattern, albeit a mystifying one. Intriguingly, when I asked MP if that is something he would say himself, he said he would, even though I have never heard or recorded him spontaneously producing such an utterance.[5]

On the documentarist side, it is interesting to recapitulate the process behind the dawning realisation that we were dealing with a real and puzzling phenomenon here by looking at the timeline of annotations, translations and analyses. For two of these texts, the ones by Copper Marnbudji and Winnie Baruwei, initial translations (into English and Kriol) were made by Sarah Cutfield (SC) and Salome Harris (SH), between 2010 and 2011, in consultation with a number of Dalabon speakers (two requested to remain anonymous in the case of the Copper Marnbudji text because parts of it are ceremonially restricted; Jimmy Wesan and Glen Wesan also contributed). Following that, NE checked the transcriptions again, partially only, with Maggie Tukumba in 2015. Then, finally, NE rechecked them with MP in 2023–2024. For the Bob Barinjbungu text, NE and MP began transcribing and translating it in February 2019 and only completed the task in

5. This is similar to the intriguing case reported by Reid (2003: 109), where modern Ngangityemerri speakers from the older generation, upon listening to a structurally very different version of their language recorded by Gerhard Laves in the 1930s, told Reid that these were structures they still recognised and were willing to repeat them back as instances of their language, even though he had never heard them himself.

2022 (usually our periods working together lasted a week and, in each day, we would spend one or two hours on transcription, interspersed with other tasks).

The earliest transcriptions of the Marnbudji and Baruwei texts, by SC and SH, convey the meaning pretty reliably, and are mostly based on a double-layered translation process, first into Kriol (the lingua franca of the region) and thence into English (translation by SC, SH, or NE), e.g. for (8) *ngahmarnûdonj* is translated in Kriol as *main dadi imin dai na-Mandolma*, lit. 'my father died, Na-Mandolma', then into standard English as 'my father died on me, na-Mandol'. Note that the "ethical dative" part of the meaning, 'on me', is introduced at the stage of English translation but is not present in the Kriol. SC and SH also introduced interlinear glosses, glossing *ngah-marnû-donj* as [1SG-BEN-die-PP], but this method of glossing does not make explicit the grammatical role of the prefix, i.e. whether 1SG is to be understood as subject, or (indirect) object, so the problematic nature of the example is not made explicit. Similarly, the translations given by MP for the Barinjbungu texts made the actual meaning, 'Hey! There's water here for us' quite clear. But of course, in all these cases the translations draw on speaker knowledge, in particular of the existence of this reversal construction. If it had just been up to the (outsider) linguist, we would have most likely adjusted what we heard to fit our preconceptions, e.g. hearing the non-reversed *kah-marnû-donj* for 'he died on me' (Example (8)) or *ngarra-h-marnû-kolh-yung* 'we put water for them' (Example (7)), adding a velar nasal to turn the intransitive verb *yu* 'lie' into the transitive verb *yung* 'put', which sounds very similar. In transcription, as in hearing your mother tongue or another language you are learning, the role of priming — hearing what you already know is possible — should never be neglected, and this is where the corrective role of native speakers and native listeners plays such a vital role.

3. Semantic nuance

One of the factors which makes for rich language documentation is careful attention to extra senses of words. We illustrate this with the verb *drahmû* (~*darhmû, dahmû*), a ditransitive verb which means 'refuse (to give to), deny to, not give to, withhold from, knock back a request'.[6] As with its antonym *won/ngabbun* 'to

6. This word is widely lexicalised in Australian languages, e.g. Bininj Kunwok *kaybun*, Kayardild *marndija* (Evans 1995: 338). In both these cases the argument projection mirrors that of the verb's antonym, to give (*won* in Bininj Kunwok, *wuuja* in Kayardild). It is tempting to relate this widespread lexicalisation to the importance of demand-sharing in these societies, and the marked and antisocial nature of refusals. Although an anonymous reviewer raises the possibility that this

give', the (non-)giver is the subject, the (non-)gifted theme is the object, and the (non-)recipient is the indirect object. And along with other ditransitive verbs in Dalabon, Bininj Kunwok (Evans 2003: 391) and other Gunwinyguan languages, the indirect object slot is realised by the "object" pronominal prefix and the theme/object slot can optionally be filled by an incorporated noun (e.g. *kah-kanj-darhminj, kah-kolh-darhminj* 'he refused me meat, he refused me drink'; here *kanj-* and *kolh-* are incorporated nominals and the prefix *kah-* means 3SG.A>1SG.(I)O). All examples in the existing Dalabon dictionary (Evans et al. 2004: 67) involve human "refusers".

Now consider the following example, involving Jack Chadum as speaker in the 1961 LaMont West recording. The context is a long description of hunting for various animals, leading to a section where they see signs of an echidna inside a hollow log, then its tracks, they look at its tracks but:

(10) *ngeh-yakkûn*[7]*-royhm-inj* *ka-h-dja-ke* *njerr*
1DIS.DU.S-R-IN.VAIN-take.a.look-PP 3SG.S-R-just-nothing 1DU.OBJ
ka-h-rdrahm-inj
3SG.A-R-withhold-PP

How should this be translated? There is no clear human agent around withholding anything. As a free translation of the whole sentence, MP suggested:

(11) 'We just looked around for nothing, but didn't have any luck at all.'

With a bit more discussion, he added that "the moon was hiding things, the moon didn't give them any good luck, this is what happens when the moon 'dies'" (sc. is no longer full). Note that there is no overt mention of the moon in any of the preceding material, and earlier parts of the passage discuss the types of hunting which are carried out in the daytime, so the whole background to this subtle translation depends on knowing the cultural script that echidnas are often hunted

---

has a more general meaning, 'someone does not want someone else to have something', I do not believe this adequately captures the meaning: On the one hand, the basic sense regularly involve active withholding, not someone simply not wanting to have something, and secondly, in the case of the moon, it requires us to personify the moon as having an intention to withhold, which is not obviously the case here.

7. Incidentally, the prefix *yakkûn-* here, another hapax, is revealing. From other speakers, a verbal prefix *wakkûn-* 'in vain' is attested, but apart from the element *-kûn*, shared with other adverbial prefixes, the form appears unmotivated. Jack Chadum's form *yakkûn*, on the other hand, is clearly related to the 'without, nothing' root *yak-* and suggests an etymological origin for the commoner (and likely innovated) *wakkûn-* form. See Caudal (2023) for an interesting discussion of how such categories, sometimes labelled "avertive" or "frustrative", give evidence for an ontological continuum between positive and negative events.

at night-time. While the precise formulation of this extra sense for the second edition of the Dalabon dictionary is still pending, and will depend on finding other examples of this type, (10) and MP's translation and explanation suggest that we should add a further idiomatic meaning, when the subject is third singular with no obvious referent and the (indirect) object is a human, of 'IOBJ have no luck (in the pursuit of something).'

This connects us back to the main argument of this paper: that linguists, to the extent that they are cultural outsiders to a large degree, will often miss out on "joining the dots" which are essential to fully comprehend particular examples. Our professional emphasis on putting in interlinear glosses to morphemes can lead us astray from the real philological quest, of working out, to the extent possible, exactly what speakers meant by the words they uttered, often in elliptical or allusive ways. Members of the speech community, especially those who are keen listeners with an extensive interpretive knowledge, are key to coming up with translations that do full justice to the intent that lay behind speakers' words, often many decades ago.

4. New vocabulary

The five recordings mentioned above — totalling just over an hour (61 minutes 50 seconds) — contain many dozens of new vocabulary items. Since these date back to a time when people regularly drew on the full panoply of Dalabon culture in their everyday lives, this provides an important resource for the evolving revisions towards a planned second and more comprehensive version of the dictionary originally published as Evans et al. (2004). I will not try to discuss them all here, but will exemplify the rich detail they involve by focussing on a single semantic domain, that of cooking vocabulary (see Table 3).

I will draw on two of the above recordings: LaMont West's recordings of Copper Marnbudji and Joy Kinslow-Harris' recordings of Bob Barinjbungu. Between them they contain fourteen previously unattested terms for methods of cooking and other food-related vocabulary. In some cases, a word is attested in both recordings, allowing us the luxury of double-checking the transcription and precise meaning.

It is worth noting that none of these terms are any longer actively used by MP. There are simply no other speakers around, at his level of fluency, to use them with, and besides that these methods of food preparation are no longer practiced. On the other hand, having grown up in the bush in very traditional conditions, he knew the food preparation methods and also recognised and understood the words at issue, from the earliest part of his lifetime. The definitions given below

attest to the detail of his knowledge. As someone with a strong commitment to maintaining his language, the chance to revive forgotten words like this is a part of the transcription process that gives him particular pleasure.

I present them below in semantic groupings, cross-referenced to the recordings, giving first the relevant Dalabon textual material and then an English translation. For conciseness, I do not use interlinear glossing in this section, but I do include comments on precise details of meaning. Where lexical words are affixed by inflectional morphology, I bold the lexical stem.

Table 3. Specific vocabulary pertaining to food preparation in the Copper Marnbudji and Bob Barinjbungu texts, which was previously unattested in investigations of Dalabon

Term	Meaning	Source
Terms for cooking large game animals, especially kangaroos		
dawaddakmû, rdawaddakmû, dawarddakmû	'cut open the belly of a kangaroo so as to put hot stones in it, to aid the roasting process while cooking in it in a ground oven' Often used in the collocation *kerri dawarddak* 'put cooking stones inside holes cut in the kangaroo' Ex: *Ken korre dahlngdawarddak!* 'Quickly put in the hot cooking stones to the holes cut in the kangaroo's belly!'	West61-Q408A4, speaker Jack Chadum, 00:02:04–00:02:05, 00:02:20–22 Bob Barinjbungu 00:24:46–48
madjarnghngun[*]	'eat small bits of offal cooked quickly on the ashes, to fend off hunger while waiting for a large game animal to cook more slowly'	West61-Q408A4, speaker Jack Chadum, 00:02:08–00:02:09
duyhbarnbun	'cover up cooking meat under the ashes'	West61-Q408A4, speaker Jack Chadum, 00:02:27–00:02:30
djirri	'juices inside roasted kangaroo' Ex.: *werrh yelûng-dedjbong barrbbarrabong djirri djohminj* 'took it out, divided it up, and sopped up the meat-juice'[**]	West61-Q408A4, speaker Jack Chadum, 00:02:35–00:02:39
djemuyirrhyirrhno	'grass in the crop of the kangaroo; when a kangaroo is killed this half-digested grass can be removed from the guts, and the kidneys and other internal parts of the kangaroo are rubbed with it to give a good flavour before cooking' In MP's variety the term *nguhdjamno* is used with the same meaning.	Bob Barinjbungu 00:23:46–49
djalerrh[***]	'get wood burning in a ground oven prior to adding cooking stones' Ex.: *Dahdjalerrh kerribard!* 'Get the wood burning for the cooking stones!'	Bob Barinjbungu 00:24:27–30

Table 3. *(continued)*

Term	Meaning	Source
ngorrowngorrow	'break the limb joints of a kangaroo so as to bundle it up for roasting' Ex.: *Ngorrowngorrow, ngarrinjdukkan!* 'Break the elbow joints then tie its hands together'	Bob Barinjbungu 00:24:30–33
yerrhbambun	'take cooked kangaroo out from ground oven' Ex.: *Ee, kah-djabuldi, woy ke ngarrahlngmadmedonj manj ngarrah-dji-yerrhbambun kuhdu ngungun kanjno* 'Yes, it's in the ground for a long time. Come on we're getting hungry, let's take it out, to eat, so we take it out of the ground oven.'	Bob Barinjbungu 00:24:55–00:25:04
minwirl	'soft juicy liquid which results from cooking the tendons of the kangaroo'	Bob Barinjbungu 00:25:04–00:25:06
daklak	'part of roasted kangaroo where the roasting juices have been sealed inside'	Bob Barinjbungu 00:25:06–00:25:07
monbubbu	'butchering cut made vertically down the front'	Bob Barinjbungu 00:25:07–00:25:08
marliyûmûno	'cut on the kangaroo's midline'	Bob Barinjbungu 00:25:09–00:25:11
Terms for getting honey		
monbu	'split tree to get native honey out' Ex.: *yeleng monbu, ngeh-monbuminj* 'then splitting the wood to get the honey out, we split the wood to get the honey'	West61-Q408A4, speaker Jack Chadum, 00:03:10–00:03:11
Terms for the food quest, hunger and abandonment		
dardawo	'give up on consuming something — e.g. food or smoke — because it's not available' Ex.: *kerna yilah-dulh-mey beka-kurnh nidjarra yalah-worrhminj djehneng yilah-wurngmi, dardawo ma yalahlng-beng-wohdanginj yeleng-ngurd* 'Later we got wood for a special fire that was supposed to be to light tobacco, but we just gave up and settled for nothing.'	Bob Barinybungu 00:16:32–40

* The corresponding noun *madjarnghno,* referring to the rapidly cooked bit of offal, was already known and listed in the Dalabon dictionary (Evans et al. 2004: 210), but the verb *madjarnghngun*, which incorporates this into the verb *ngun* eat, was not known at the time of making the dictionary.
** Listening to the text, MP further noted that a type of plant called *munmun* was usually used to soak up these juices.
*** Since there is a prefix *dja-* 'just', this may simply be *lerrh*, though that is not attested in the dictionary; more probing needed.

Once again, working with someone like MP, with his extensive receptive knowledge of Dalabon and his eagerness to listen to old recordings, allows us to significantly deepen our vocabulary. This will be incorporated into the next and expanded version of the Dalabon dictionary, greatly enriching our lexical coverage. It is not that MP isn't a fluent speaker — he is — but the lack of other fluent speakers to talk to, coupled with the disappearance of many traditional practices decades ago, means that these words were not in his active vocabulary, nor, for that matter, in the active vocabulary of other speakers close to a generation older, who I worked with in the two decades since the mid 1990s. I would also like to emphasise that this is not simply a question of deciding what a transcribed but unfamiliar word means. Listening to words spoken at fluent speed, without the possibility of repeating them and being corrected by the speaker, is frequently insufficient to come up with an accurate transcription. MP, on the other hand, would be quite clear about the correct pronunciation — in effect (and unlike a linguist) he is not listening to just one token of a word, but matching it against a word he had heard sufficiently often in the past to have built up a clear phonological representation.

5. Context, and interpretation: The case of triangular kin terms

Dalabon, like many other Australian languages, has a number of "triangular" (or "trirelational") kin terms, which simultaneously express the kinship relation between speaker and referent, and addressee and referent, in the context of a particular presupposed relation between speaker and addressee (Evans 2021: 73–74). For example, if I use the term *kundjirr* when talking with my sister- or brother-in-law Jo, this could refer either to my sibling or Jo's sibling; it thus means 'one who is sibling to one of us and sibling-in-law to the other of us, given that you and I are brother/sister-in-laws'. Likewise, if I am talking to my same-sex sibling Alex (i.e. my brother Alex if I am a man, or my sister Alex if I am a woman) and I use the word *ngalkeywurd*, this means 'the daughter of one us and same-sex sibling's daughter of the other, us two being same-sex siblings'. (In Dalabon society, as in other Australian Aboriginal societies, kinship is "universal" in the sense that kinship relations are extended out to everyone in the social universe, bar outsiders like government officials, and part of being socialised as a Dalabon speaker involves knowing everyone's relationship to everyone else.).

In Dalabon, these triangular terms are known by the metalinguistic term *derbuy* (sometimes pronounced *drebuy*), and the derived verb *drebuykan* (*kan* is etymologically 'carry') means 'to use indirect language in reference to kin'. In fact, the word *derbuy* has a number of semantic layers: As well as the specific meaning 'triangular kinship vocabulary', it can also mean 'polite, delicately-worded talk'. It can also mean 'subtlety' in a more general sense, as in the preface to the 2004

Dalabon dictionary, for which Maggie Tukumba dictated her wish that through the dictionary *kenbo yilah-dulu-burlhkeywoyan, mak ka-dulu-wanjingh, bah ka-djahlng-ngong-no kanh dulu-no, kanh drebuy njelng yilaye-yenjdjung* 'then we'll bring out all our words, not just one word, but all sorts of words, including the subtleties of what we say' (Evans et al. 2004:vii).[8]

Now our best understanding of how this phenomenon, found in around 15 languages of Australia (McConvell 2023), and Mebengôkre in the Brazilian Amazon (Lea 2004; Coelho de Souza 2020), comes from studies that show how these terms are employed in actual conversation for purposes such as footing. For Australian languages, Blythe (2018) and Garde (2013) are the most notable examples in this tradition; importantly, both could draw on rich situated evidence from conversations held in languages (respectively Murrinhpatha and Bininj Kunwok) that are still vibrant, being learned by children, and offer many opportunities for observing conversation between adults who still deploy the full panoply of their stylistic resources in everyday life.

Sadly, this is no longer the case for Dalabon. Although there has been substantial work in the last two and a half decades by a range of researchers — particularly Sarah Cutfield, Francesca Merlan, Maïa Ponsonnet and myself — this has been in settings where Dalabon is no longer the language of daily use, having ceded this role to Kriol, Mayali (a variety of Bininj Kunwok) and English. Just four triangular kin-terms (Table 4) were included in the only published dictionary of Dalabon so far (Evans et al. 2004), elicited by such procedures as asking for the Dalabon equivalents of triangular kin-terms in Bininj Kunwok (which most Dalabon speakers also know). It is revealing that all the examples in the dictionary are generalised explanations of how these terms are used, rather than actual examples of their use in context, e.g. *Kundjirr-no kanh yale-yin kirdikird-no ka-kan yilah-derbuykan, mak yila-ngeybun wulkûn-njelng. Mak yila-ngeybun wulkûn-njelng.* 'When we say *kundjirr*, he's got his wife, we refer to him indirectly, we can't call the names of our brothers.' (Evans et al. 2004: 184).

As mentioned above, however, we are fortunate to have Dalabon recordings from earlier periods of time when the language was in everyday use by speakers maintaining a wide stylistic range. Of the five recordings mentioned above, the most important one here is the one made by Barry Alpher in 1974 at the settlement of Bamyili, assisted by Dalabon heritage speaker David Jentian Nangan.golod

8. The root *dulu*, also occurring here, normally occurs in its 3rd person possessed form *duluno*, and has a range of meanings including 'word', 'rule, law', 'song (word), dreaming passage of song or ceremony'. The interesting use of *drebuy* here, in the sense of 'subtlety', is comparable to the French *tournure* or English *turn of phrase*, a pivoting of reference (first to kin, then to other meanings) by changing the perspective.

Table 4. Triangular kin terms (*derbuy*) contained in the Dalabon dictionary (Evans et al. 2004)

Term	Meaning
kundjirr	'polite way of referring to one who is sibling to one of us and sibling-in-law to the other of us, given that you and I are brother/sister-in-laws, e.g. a woman would use this term with her brother's wife when referring to her brother'
nakeywurd	'polite way of talking, between siblings, to refer to the son of one of them'; 'son of one of us; sibling's son to the other'
yengkuldjanngu	'polite way of referring to someone who is the niece of one and the daughter of the other'; 'niece to one of us, daughter to the other, we being cousins'
yengkulngu	'polite way of referring to someone who is the nephew of one of the interlocutors and the son of the other'; 'nephew to one of us, son to the other, we being cousins'

who was at that time a student at the School of Australian Linguistics in Batchelor, Northern Territory, where Alpher was employed as a lecturer.

Alpher was particularly interested in investigating the two sets of dual pronominal prefixes, e.g. *barra-* vs *ke-* for 3rd person dual, which canonically distinguish "harmonic" (= even-numbered generations like siblings or grandparent/grandchild) from "disharmonic" (= odd-numbered generations like parent/child), and this recording forms the basis for his influential article (Alpher 1982) on this pronominal series. We note in passing that already in that article he mentioned the possibility of extending the "disharmonic" use of pronouns to other contrast pairs, such as an Aboriginal person and a white person (Alpher 1982: 21, ex. 7); this usage, which he recorded in 1972, is confirmed by an example involving another speaker, Lofty Nabardayal Nadjimerrek, recorded ten years earlier by Joy Kinslow-Harris at Oenpelli, where it is exemplified by the prefix *ke-* in the sentence *kehboniyan wahdû, nûnda Midjidj kehyongiyan wahdû* 'the two of them will go, including the white person, that white woman, they will camp together' (KINSLOW-HARRIS_J03-000320A.TMP: 54:50–54:54).

The recording features a number of Dalabon speakers — Winnie Baruwei (WB), David Jentian (DJ), Jack Chadum (JC) and others who come and go during the recording. In terms of the kinship relations obtaining between those in the conversation, WB calls JC 'right skin', i.e. eligible marriage partner, hence *duwayngan* 'my husband', JC and DJ count as brothers to each other despite having different "skins" or subsections (Bulanj and Ngarridj respectively)[9] It is a mix of elicited

9. Within the intricate system of marriage rules regulated by eight "skins" or subsections (see Evans 2021: 42), there are four pairs of skins which are equivalent for many purposes of kin-

material[10] and more chaotic, naturalistic conversation, which turns out to be the most revealing material for the purposes of understanding triangular kin term use. It contains four *drebuy* terms — *kundjirrngu, djumungu, yawnjung*, and *nabbolo* — of which only the first was included in the Dalabon dictionary (Evans et al. 2004). That dictionary, deliberately entitled *A First Dictionary of Dalabon*, with around 4,000 entries, only included four *drebuy* terms of which the last two are simply masculine and feminine variants (see Table 4 above). Compare this to well over 100 terms in closely related Bininj Kunwok (Garde 2013). The short Baruwei text thus nearly doubles that number.

We will discuss just one such case here, namely the term *yawnjung,* to illustrate the referential intricacies involved. The relevant passage of the Alpher 1974 recording, spoken by Winnie Baruwei, runs (00:13:57–00:13:59):

(12) a. *wo, yawnjung-yih*
yes brother.of.one.of.us.brother.in.law.of.one.of.us-INST
njehka-h-ma-nginj
3SG.A>12DU.O-R-get-PP
'Yes, he recorded you and me, the one who is brother of one of us and brother-in-law of the other.'

Now it is perfectly possible to gloss this and to translate it in the way done above. But that is not the same as knowing what it means in context, in particular, who is being referred to by this expression. She could have used a more general term meaning 'brother', but given the "classificatory" nature of Dalabon kinship, which extends kinship relations out to the entire social universe, this would still leave an enormous number of potential referents. By using the term *yawnjung* she adds the further referential specification that, taking into account the kin relationships between her, her addressee, and the referent, she is talking about someone who is brother-in-law of one of them and the brother of the other. We already know that Jack Chadum is the addressee: This is inferable from the fact that he is in the immediately preceding turn. To that, MP added his knowledge that Jack Chadum was WB's brother-in-law, and that David Jentian Nangan.golod (DJ) was Jack's brother and her brother-in-law. From this, with some careful thought as we listened through the recording a couple of times, MP was able to infer that she is talking about DJ. We can therefore recast the English translation as (12b), with a precise named referent.

---

reckoning, since they can be assigned based on marriage of the same parent to two different spouses, one drawn from the first-choice spouse category and the other from the second. Bulanj and Ngarridj are one such pair.

10. Starting with elicitations of dyadic kin terms and kinship-sensitive pronominal prefixes, material that appeared in Alpher (1982).

(12) b. *wo, yawnjung-yih*
yes brother.of.one.of.us.brother.in.law.of.one.of.us-INST
njehka-h-ma-nginj
3SG.A>12DU.O-R-get-PP
'Yes, David Jentian Nangan.golod recorded the two of us.'

This sort of cryptic-crossword decipherment, which is a regular part of how speakers of Australian Aboriginal languages establish reference, could be carried out by MP in listening to the recording, because he knows the kinship relationships of all involved, and because he is used to conversations in which reference is achieved by these means. This sort of hermeneutic enrichment, half a century after the original recording and by arguably the last fluent speaker, exemplifies how great the difference is, in utility and interest, between a basic recording, or even one with a first-pass transcript, and a careful transcript and translation bringing the full panoply of speaker-intuition to the problem of achieving 'person reference in interaction' (Enfield & Stivers 2007).

6. Concluding remarks

The products of the last three decades of research inspired by the documentary-linguistics program have plumbed a rich and far-reaching sample of the world's languages, along with the cultures they are embedded in and which they enable people to talk about. Rapid advances in recording, analysis and archiving methods have begun to make many more wings of the great "Library of Babel" available to those whose curiosity takes them there, and especially for those in the future who will look for words, ideas, meanings, stories, songs and bodies of knowledge from fragile oral traditions which, as one of the great extinction tragedies of our times, are likely to disappear or at least lose much of their rich detail.

Yet without careful attention to fine problems of translation, these recordings will often fail to realise much of their potential, as a record of the language. In this paper we have examined four manifestations of this problem, with regard to the Dalabon language. In each case Manuel Pamkal's deep receptive knowledge of his language enables a special type of "dialogue with the ancestors" — in particular, here, a "dialogue with the archives". Coupled with his motivation and care in listening over old recordings, his considered and precise translations have turned recordings which would have at best been semi-interpretable, or replays of what we already believed about the language, into treasure-troves of valuable and surprising new material.

In Section 2 we saw how these old recordings contain interesting examples of "argument reversals" revealing the opposite mapping to what we would expect,

as between indexing prefixes and the subject and indirect object arguments. In Section 3 we showed how cultural knowledge, allowing the restitution of key parts of what had been left unsaid by the speaker, allows us to identify an additional and interesting meaning for a key lexical item. In Section 4 we showed how two texts recorded in the 1960s abound in food-preparation terms, pertaining to now-abandoned cultural practices, that had simply stopped appearing in texts recorded from the 1990s on, but which MP was able to interpret thanks to hearing about them in his youth. And in Section 5 we saw how, thanks to the earliest recording of actual Dalabon conversation, we were able to find examples of triangular kinship terms in actual use (rather than prescriptive statements about how they should be used), and combine these with MP's knowledge of the kinship relationships between the speakers to recapitulate the inferential steps used by speakers to reach beyond the terms' general meaning and work out their actual referents.

Notwithstanding Elon Musk's[11] recent hyperbolic proclamation that "[t]he cumulative sum of human knowledge has been exhausted in AI training. That happened basically last year" the majority of languages, and the full richness of the cultures they transmit, remain unknown to science (Himmelmann 2014, Evans 2022, Skirgård et al. 2023). New phenomena are constantly turning up as the field gradually extends its coverage in the race against language death in many parts of the world (Seifart et al. 2018). And one of the great barriers to discovering these new phenomena comes from the processes of extrapolation and interpolation — using what we know, to project how we think new data (which might in fact be heritage data like old recordings) should be fitted into the framework of what we know. Whether we look at AI training, or running interlinear glossing programs to determine how words in new text should be parsed, we face the same risk. And it is by working with speakers who listen closely for what is meant that we are best able to avert it and make the new discoveries we seek.

Exploring such issues to the full remains a challenge that must be approached from many angles. At the heart of this is the question of how to understand the bidirectional mapping between form and meaning in each language, mediated by grammar, and situated in contexts of discourse, pragmatics, cultural context, psycholinguistic framing, sociolinguistic placement, historical trajectories, gestural elaboration, and much else besides. For the reasons given above, that central mapping between form and meaning will always be asymmetric within the great project of documentation. We can record form directly, but must understand meaning through a range of more indirect processes including contexts, paraphrases and translations, well-designed stimulus sets linked to the utterance, uptakes by other

11. https://www.theguardian.com/technology/2025/jan/09/elon-musk-data-ai-training-artificial-intelligence

recorded speakers, and participant observation. Inevitably, this means that our documentation of the form side goes faster, and is more amenable to measure and analysis by objective means. But without due attention to the meaning side, we risk problems like those described below, in another publication emanating from the Cologne linguistic tradition, right at the beginning of the documentarist era:

> The worst case — found all too often — is an immaculate sound recording of a passage in language, without translation. For a language about which little is known this is about as helpful as tablets in the Indus or other undeciphered scripts: we recognize that language is there, without knowing what it means. Such cases can result either from language materials that are recorded without being analysed, or through a prevalent asymmetry by which the original text is recorded but not the process of arriving at a translation through subsequent discussion and probing. There are also a range of other less than perfect outcomes, such as translations which are wrong, or too specific (e.g. leaving out alternative translations), or simply uncheckable because it is impossible for subsequent investigators to go back and work out why the given translation was arrived at.
>
> (Evans & Sasse 2007: 65–66)

Fortunately, though, for at least some speakers the ability to understand and comment sensitively upon what is meant persists for many decades after the diminishment of a language as a vigorous everyday vehicle of communication. In this article we have sought to illustrate how it is possible to draw on this knowledge, in revisiting older language documentations, in ways that partially rectify the form-meaning asymmetry and which, through annotations which may be made more than half a century after the original recording, begin to supply some of the interpretive keys needed to draw closer to the meaningful heart of speech.

The cases discussed here are not a panacea for all problems of semantic interpretation. In the current case we were very fortunate to be able to draw on MP's exceptional talents as a linguistic bridge: growing up in the bush, in a setting where Dalabon and Bininj Kunwok were the main languages of everyday communication and many traditional practices were still being followed, laid down an experiential and linguistic expertise onto which his later learning of English (starting from his late teens) and experience in explaining his culture as a tour guide were then grafted, to make him an exceptional interpreter/translator not just between Dalabon and English but between past and present. It is by no means guaranteed that language documenters will find someone of this linguistic calibre, and certainly in my own experience I have worked with several languages (e.g. Kayardild, Ilgar) where no equivalent person could be found.[12]

12. Though see, for Marrku, the interesting case of Joy Williams who was able to help transcribe and translate a text recorded from her mother in the 1960s, even though she herself could not

But though he was able to help with many words and passages, there were many others which we were either unable to even hear clearly (e.g. we could not make out the verb stem in *yilah-kurnh-XXX* at 11.00 in the Bob Barinjbungu text), and others which we could hear but could not understand. For example, right after (9) in the same Bob Barinjbungu text, the word *yalah-yang-yininj warrahman* appears: the word *warrahman* can be heard clearly, but MP was adamant he does not know what it means. Slightly less unclear is at 5:29 in the same text, where a word sounding like *Yaladjdja* appears, right after *nabokala yala-kolhnang* 'we saw the saltwater'; here MP thinks it might mean 'saltwater', or the name of a place on the sea, but is not sure. Without MP's knowledge of Dalabon, there would have been many more examples like this, and they are typical of the sorts of problems that would face any attempt to transcribe and translate texts that have been recorded but not translated.

In the case of the texts treated here, all recorded as side-activities in the 1960s and 1970s by researchers primarily concerned with other agendas,[13] we are simply grateful that they exist at all: These recordings predate other material on the language by two to three decades. Not only do they enrich what we know about the lexicon and grammar of Dalabon, and give us an invaluable temporal reach in building a corpus able to trace actual change through time, but they also enable speakers like MP to integrate language documentation with language revival[14] through the potential for ongoing "dialogues with the ancestors" that this process has afforded.

But what lessons should documenters draw when they are working now, in the documentarist paradigm, with an eye to the future? Of course, it will never be possible to transcribe and translate everything that is recorded. What is possible, though, is a redistribution of effort so that fine and situated translations, and careful explanations of the meanings of words and phrases, are allocated a

---

speak Marrku. In that case, though, her knowledge was much less far-reaching than MP's, and it was very much confined to transcriptions of her own mother's speech, whereas she found it difficult to make much sense of recordings made with others. See Evans (2001, 2022: 210–211).

13. Joy Kinslow Harris was gathering brief comparative materials for a thesis on the internal structure of the Gunwinyguan family (Harris 1969), with her primary focus being on the related language Kunbarlang (Gunbarlang), LaMont West was engaged in primarily anthropological research, and Barry Alpher was employed at the School of Australian Linguistics in Batchelor, Northern Territory, where David Jentian Nangan.golod was a student, and was engaged in a short and focussed fieldtrip with the latter, to figure out the use of the "harmonic" and "disharmonic" pronominal prefixes, published as Alpher (1982).

14. Cf. Ramallo and Farfán, on "allowing a more equal participation on the side of the speaker in the whole process [of language documentation]" (2010: 6), and of "identifying and analyzing speakers' expertise knowledge as manifested in heterogeneity of players and actors" (2010: 10).

fair share of the project's time and attention. Four other strategies that are likely to improve the quality of transcription and translation are the careful inclusion of context through video (but we should not exaggerate here: This is only really useful when what is being talked about is also being demonstrated), the embroidering of hermeneutic discussions into the documentarist enterprise (i.e. treating discussions of translation and meaning as worthy of being counted in the number of hours targeted in a documentation project), the involvement of younger speech community members who may not appear to be fully fluent active speakers but may have exceptional skills as interpreters and semantic commentators, and a commitment to the long term, so as to allow the careful revisiting, retranscription and translation or re-translation of previously recorded materials.

As Nikolaus Himmelmann has pointed out many times, it is vital that language documentation put itself squarely in the shoes of future users of archived language materials. In this article we have tried to give a feeling for one way of doing this, by working from the perspective of present, to recordings made in the broad spirit of language documentation, but long before the term itself was coined.

Funding

I would like to acknowledge the support of these institutions and projects for my work on Dalabon: Data on Dalabon has been gathered over many years and many grants, with financial support from the Australian National University, Maningrida Arts and Crafts, Mimal Land Management, the Hans Rausing Endangered Languages Project, the Alexander von Humboldt-Stiftung (Anneliese Maier-Forschungspreis), and especially the Australian Research Council through a series of projects: Non-Pama-Nyungan Languages of Northern Australia: descriptive, grammatical, comparative and sociolinguistic investigation, Reciprocals Across Languages, and most recently Language and Social Cognition: The Design Resources of Grammatical Diversity (DP0878126), as well as the ARC Laureate Project The Wellsprings of Linguistic Diversity and supplementary support from the ARC Research Centre for the Dynamics of Language (CoEDL) (CE140100041).

Acknowledgements

I am grateful to Birgit Hellwig and to two anonymous referees for their helpful comments on an earlier version of this article, to Sarah Cutfield and Salome Harris for their earlier work wrestling with first transcriptions of this material, to Murray Garde, Francesca Merlan and Maïa Ponsonnet for long-running discussion on cultural details of languages of the region, and to the many Dalabon speakers, sadly most deceased, for their generous teaching of their language over many years. Finally, I would like to express my thanks to my many friends and colleagues at the Institut für Linguistik at the Universität zu Köln for many discussions of language documentation and semantic fieldwork over the years.

Author statement

A note on authorship: this article was written wholly by Nicholas Evans, but because of Manuel Pamkal's active and far-reaching involvement in the transcription tasks discussed here, and in the whole research project more generally, it is appropriate to include him as a co-author. Through the text, however, when the first person singular is used, it refers to Evans, since it was Evans who actually wrote the prose of the article.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following glosses and abbreviations are used:

>	acting upon (subject acting upon object or indirect object)
৸	reversal reading
AIATSIS	Australian Institute of Aboriginal and Torres Strait Island Studies (Archive)
DIS	disharmonic
(1)O	object/indirect object (distinguishable by other grammatical tests but not by indexation)
IN.VAIN	in vain
INST	instrumental
ITER	iterative
PART	particle
PP	past perfective
R	realis
RR	reflexive/reciprocal
TIME	'time when'.

References

Alpher, Barry. 1982. "Dalabon Dual-subject Prefixes, Kinship Categories and Generation Skewing." In *Languages of Kinship in Aboriginal Australia*, ed. by Jeffrey Heath, Francesca Merlan, and Alan Rumsey, 19–30. Sydney: Oceania Linguistic Monographs.

Baerman, Matthew. 2007. "Morphological Reversals." *Journal of Linguistics* 43 (1): 33–61.

Blythe, Joe. 2018. "Genesis of the Trinity: The Convergent Evolution of Trirelational Kinterms." In *Skin, Kin and Clan: The Dynamics of Social Categories in Indigenous Australia*, ed. by Patrick McConvell, Piers Kelly, and Sébastien Lacrampe, 431–471. Canberra: ANU Press.

Capell, Arthur. 1962. *Some Linguistic Types in Australia*. (Handbook of Australian Languages, Part 2.) Sydney: University of Sydney.

Caudal, Patrick. 2023. "Avertive/Frustrative Markers in Australian Languages: Blurring the Boundaries between Aspectuo-temporal and Modal Meanings." In *Understanding Human Time*, ed. by Kasia M. Jaszczolt, 107–173. Oxford: Oxford University Press.

Coelho de Souza, Marcela S. 2020. "Porque o parentesco é sempre triádico." In *Macro-Jê: Saberes Entrecruzados*, ed. by Maxwell Miranda, Águeda Aparecida da Cruz Borges, Áurea Cavalcante Santana, and Suseile Andrade Sousa, 193–220. Barra do Garças: Ceddeli.

Cutfield, Sarah. 2011. Demonstratives in Dalabon, a Language of Southwestern Arnhem Land. PhD thesis, Monash University.

Durantin, Gautier, Ben Foley, Nicholas Evans, and Janet Wiles. 2017. "Transcription survey." Unpublished survey, COEDL (Australian Research Council Centre of Excellence for the Dynamics of Language).

doi Enfield, Nick, and Tania Stivers (eds). 2007. *Person Reference in Interaction: Linguistic, Cultural, and Social Perspectives*. Cambridge: Cambridge University Press.

doi Evans, Nicholas. 1995. *A Grammar of Kayardild*. Berlin: De Gruyter Mouton.

doi Evans, Nicholas. 2001. "The last speaker is dead — long live the last speaker!" In *Linguistic Field Work*, ed. by Paul Newman, and Martha Ratliff, 250–281. Cambridge: Cambridge University Press.

Evans, Nicholas. 2003. *Bininj Gun-wok: A Pan-dialectal Grammar of Mayali, Kunwinjku and Kune*. Canberra: Pacific Linguistics. handle:http://hdl.handle.net/1885/53188

doi Evans, Nicholas. 2013. "Instability in Stability: Therapeutic Periphrasis in the Dalabon Verbal Prefix Paradigm." *Proceedings of the British Academy* 180: 53–76.

doi Evans, Nicholas. 2021. "Social Cognition in Dalabon." In *Language Documentation and Conservation SP 12. Social Cognition Parallax Interview Corpus (SCOPIC)*: 22–84.

Evans, Nicholas. 2022. *Words of Wonder: What Endangered languages Tell Us*, 2nd ed. Maldon and Oxford: Wiley Blackwell.

doi Evans, Nicholas, and Alexandra Marley. 2023. "The Gunwinyguan Languages." In *The Oxford Guide to Australian Languages*, ed. by Claire Bowern, 781–795. Oxford: Oxford University Press.

doi Evans, Nicholas, and Hans-Jürgen Sasse. 2007. "Searching for Meaning in the Library of Babel: Field Semantics and Problems of Digital Archiving." *Language Documentation and Description* 4: 58–99.

doi Evans, Nicholas, Dunstan Brown, and Greville Corbett. 2001. "Dalabon Pronominal Prefixes and the Typology of Syncretism: A Network Morphology Analysis." In *Yearbook of Morphology 2000*, ed. by Geert Booij, and Jaap van Marle, 187–231. Dordrecht: Springer.

doi Evans, Nicholas, Janet Fletcher, and Belinda Ross. 2008. "Big Words, Small Phrases: Mismatches between Pause Units and the Polysynthetic Word in Dalabon." *Linguistics* 46 (1): 87–127.

Evans, Nicholas, Francesca Merlan, and Maggie Tukumba. 2004. *A First Dictionary of Dalabon (Ngalkbon)*. Maningrida: Bawinanga Aboriginal Corporation.

doi Foley, Benjamin, Josh Arnold, Rolando Coto-Solano et al. 2018. "Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (Elpis)." Presentation, 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages, Gurugram, 29–31 August 2018.

doi Garde, Murray. 2013. *Culture, Interaction and Person Reference in an Australian Language: An Ethnography of Bininj Gunwok Communication*. Amsterdam: John Benjamins.

doi Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel (eds). 2006. *Essentials of Language Documentation*. Berlin and New York: De Gruyter Mouton.

Harris, Joy Kinslow. 1969. Descriptive and Comparative Study of Gunwinyguan Languages, Northern Territory. PhD thesis, Australian National University.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2006. "Language Documentation: What is it, and what is it good for?" In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin and New York: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2014. "Sprachenvielfalt als natürliches Experiment." *Spektrum der Wissenschaft* 5: 68–71.

Lea, Vanessa. 2004. "Aguçando o entendimento dos termos triádicos Mebengôkre via aborígenes australianos: dialogando com Merlan e outros." *LIAMES* 4: 29–42.

Marley, Alexandra. 2020. Kundangkudjikaberrk: Language Variation and Change in Bininj Kunwok, a Gunwinyguan Language of Northern Australia. PhD thesis, Australian National University.

McConvell, Patrick. 2023. "Kinship, Marriage and Skins." In *The Oxford Guide to Australian Languages*, ed. by Claire Bowern, 520–529. Oxford: Oxford University Press.

Ponsonnet, Maïa. 2013. *A Culturally Informed Corpus of Dalabon*. ELAR. handle:http://hdl.handle.net/2196/00-0000-0000-0008-5E9C-6

Ponsonnet, Maïa. 2014. *The Language of Emotions. The Case of Dalabon (Australia)*. Amsterdam: John Benjamins.

Ramallo, Fernando, and Jose Antonio Flores Farfán. 2010. "Exploring Links between Documentation, Sociolinguistics and Language Documentation." In *New Perspectives on Endangered Languages*, ed. by Jose Antonio Flores Farfán, and Fernando Ramallo, 1–12. Amsterdam: John Benjamins.

Reid, Nicholas. 2003. "Phrasal Verb to Synthetic Verb: Recorded Morphosyntactic Change in Ngan'gityemerri." In *The Non-Pama-Nyungan Languages of Northern Australia. Comparative Studies of the Continent's Most Linguistically Complex Region*, ed. by Nicholas Evans, 95–123. Canberra: Pacific Linguistics.

Seifart, Frank, Nicholas Evans, Harald Hammarström, and Stephen Levinson. 2018. "Language Documentation 25 Years On." *Language* 94 (4): e324–e345.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi et al. 2023. "Grambank Reveals the Importance of Genealogical Constraints on Linguistic Diversity and Highlights the Impact of Language Loss." *Science Advances* 9 (16).

CHAPTER 6

Language archives in context

Felix Rau
University of Cologne

Language archives are a type of research infrastructure that grew out of requirements and efforts of language documentation. From the very beginning of language documentation as a separate research paradigm, it has been accompanied by a strong and vibrant methodological discourse. This discussion of methods and data has not only led to a clear framework of documentary linguistics, but also specified requirements for data and metadata formats, software tools, language archives and ontologies. A whole ecosystem of data infrastructure along with new data types has grown out of this methodological discourse and in close dialogue with language documentation practice.

Keywords: research data, research infrastructure, methodology, linguistic evidence

1. Introduction

Language archives are institutions that preserve digital audio-visual recordings of speech events and make them available to researchers, the speech community, and the general public. Digital language archives as we know them today have developed in parallel to the formation of documentary linguistics as a sub-field of linguistics. In this paper, we will take a closer look at this relationship. We will emphasize the fact that language archives are a type of research infrastructure that grew out of the conceptual and methodological requirements of language documentation. However, language archives are also part of a larger landscape of research infrastructures for language data. As institutions in this landscape, they have a wide variety of stakeholders that include speakers/signers, speech communities, researchers, the archive's hosting institution, and funding agencies. All these relationships have formed language archives as institutions and continue to do so.

This contribution will not offer another comprehensive history of language archiving. Henke and Berez-Kroeker (2016) and Berez-Kroeker and Henke (2018) present a detailed history of language archiving. Henke and Berez-Kroeker (2016)

https://doi.org/10.1075/slcs.240.06rau
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

provide a narrative based on four periods based on the changes in focal topics and technical developments. They identify a first *early period* that lasted from the beginning of philological or linguistic language description until the beginning of the 1990s. The *second phase* starts with the recognition of language endangerment and language documentation and lasts for about fifteen years until 2006. Overlapping with the second period, a *third period* lasted from 2000 to 2010 and was centered around technological advancements, standardisation, and largescale funding efforts. The *fourth period* started 2010 and lasted at least until the paper was published in 2016 and this phase is characterized by participatory models of research. While this division into phases makes a good narrative, and is probably necessary when placing developments in a historical context, it deemphasizes the continuous dialogue between methodology and infrastructure, the asynchronous developments in different parts of the world, and academic and institutional context in which languages archives operate and develop.

The account from Henke and Berez-Kroeker (2016) and Berez-Kroeker and Henke (2018) gives a thorough overview and is very illuminative. However, in our paper we want to take a complementary perspective and focus on the continuous relation between the methodological discourse of documentary linguistics and language archives. As a starting point, we note that most — if not all — central conceptual and methodological issues of the development of language archives are already touched upon in Himmelmann (1998). In the past quarter century, the methodological discourse around language documentation has obviously continued to develop, change and deepen, but the discourse has mostly elaborated on topics set by Himmelmann (1998). This highlights not only the seminal character of Himmelmann (1998), but also the continuous back and forth between methodology and infrastructure. And while technical developments have facilitated developments, it was the theoretical and methodological discourse that took primacy and that defined the scope for such developments.

From the very beginning of language documentation as a separate research paradigm, it has been accompanied by a strong and vibrant methodological discourse (see e.g. Himmelmann 1998, 2006, 2012; Woodbury 2003; Lehmann 2001; Austin 2013; Seifart et al. 2012, 2018; Thieberger 2016). This discussion of methods and data has not only led to a clear framework of documentary linguistics, but also specified requirements for data and metadata formats, software tools, access protocols and (meta-)data ontologies. A whole ecosystem of data infrastructure along with new data types has grown out of this methodological discourse and in close dialogue with language documentation practice (Seifart et al. 2012; Thieberger 2016; Himmelmann 2018; Rice & Thieberger 2018). We argue that the relationship between methodological discourse and language archives shaped this type of institution and that in fact research infrastructure is at its best when it is

motivated by immediate requirements from research methods and when developments in research infrastructure are directly tied to developments in the methodological discourse of the field.

2. The emergence of documentary linguistics

The seminal paper for language documentation as a separate field of linguistics is Himmelmann (1998). While documentary practices have been part of field linguistics for a long time, Himmelmann (1998) established documentary linguistics as an independent paradigm. He argued for the separation of documentary linguistics from descriptive linguistics based on its distinctive methods. Himmelmann (1998) touches on most methodological aspects that become relevant in the following decades for documentary linguistics and the developments of language archives. This includes corpus design, corpus structure, data quality, data ownership and copyright, privacy rights, community participation, sustainable archiving, access control.

Henke and Berez-Kroeker (2016: 413–415) argue that language archives existed before the advent of documentary linguistics as a separate field. Field linguists deposited their field notes and occasionally analogue recordings on various media in appropriate institutions, in particular ethnographic or anthropological archives or museums. However, it has been claimed that this practice mostly ceased after the middle of the twentieth century (Johnson 2004: 141), maybe because of the rise of new paradigms in linguistic theory (Henke & Berez-Kroeker 2016: 415). In the pre-1990s language archives as described by Henke and Berez-Kroeker (2016) and Johnson (2004), the focus was on physically preserving by-products of language descriptions. This focus resulted in an emphasis on preservation of physical artifacts of research, less on access to — let alone publication of — the data. As the materials deposited and preserved were dominantly by-products of language description, there was also no methodological discourse about the structure, content, or quality of these deposits. We would therefore claim that the character and role of these archives is decidedly different from current language archives, even though some language archives grew out of this type of earlier institutions.

Himmelmann (1998) constitutes a considerable change for the methodological status of language archives. He identifies "presentation for public consumption/publicly accessible storage (archiving)" (Himmelmann 1998: 171) as one of four steps of language documentation. The aspect of access and the methodological as well as social and ethical consequences of making primary data accessible are central to documentary linguistics as envisioned by Himmelmann (1998, especially Section 3.1). Furthermore, Himmelmann (1998) addresses the issues of corpus

design (Section 3.2) and data quality (Section 3.3). The methodological discourse in documentary linguistics was indeed decidedly different from descriptive linguistics and this difference shaped the development of language archives after 1998. In the following years, developments in the methodological discussions about documentary linguistics would accompany changes to language archives and the practices surrounding these institutions. In Section 3, we will look at technical developments. Section 4 examines data quality and developments in the criteria for corpus creation. In Section 5, the changing place of language archives in the broader research infrastructure landscape will be explored. In the following Section 6, we will broaden the view even more and discuss the place of language archives in a wide network of stakeholders. An important area of methodological discussion and an area where language archives display special competencies decidedly different from many other parts of the research infrastructure landscape.

3. The technical developments and method

Language archives as repositories for digital audio-visual corpora of (originally mostly) endangered languages have — from their beginning — been institutions with a strong and defining technical side. In contrast to Henke and Berez-Kroeker (2016: 416) who focus on the aspect that "modern documentary linguistics was able to emerge due to monumental developments in digital information technology", we want to emphasize that there is no primacy of technology. Methodological goals can often be achieved in various technical ways, and while there is a dialectical relationship between methodological discussions and current technical possibilities, implementation does not automatically follow technical developments.

The focus on raw and primary data in Himmelmann (1998: 161) was motivated by a methodological emphasis on primary data when working with endangered languages. Recording and transcribing audio (or even audiovisual) data had been part of the research methods of various linguistic subdisciplines including descriptive linguistics, conversational analysis, discourse analysis and phonetics since the advent of the phonograph cylinder (see for example McGregor 2008: 419–425). The centrality of primary data in documentary linguistics and the concern with producing and preserving a record of the linguistic practices of a speech community encouraged a deeper engagement with the technical possibilities in the second half of the 1990s and in particular in the early years of the 2000s. This included the development of ELAN (Wittenburg et al. 2002, 2006; Sloetjes & Wittenburg 2008), a tool for adding time-aligned annotations to one or more media files. It is apparent from the early discussions of endangered languages

beginning with Johannes Bechert's presentation in 1987 (published as Bechert 1990) and followed by Hale et al. (1992) as well as from early works on documentary linguistics, that the increased availability of digital recording devices and an increasing prominence of multi-media records in computing did not initialize the field of documentary linguistics or its focus on primary data. However, the concrete implementation of ELAN as a Java application creating and manipulating an application-specific XML-based annotation file separate from the audio or audiovisual media file(s) is clearly due to technological developments in the early 2000s.

The centrality of preservation and publication of primary data has not only led to new forms of digital data representations, it has also initiated an intense discourse on data documentation and metadata. From very early on, language documentation and language archives have addressed issues of data description with rich metadata, metadata publication and access control (e.g. Broeder et al. 2000; Bird & Simons 2001; Broeder et al. 2001; Nathan & Austin 2004; Broeder & Wittenburg 2006; Trilsbeek & Wittenburg 2006; Sullivant 2020). After its foundation in 2003, the Digital Endangered Languages and Musics Archives Network (DELAMAN) became a forum for technical coordination and standardization among language archives.[1] The OLAC metadata standard is based on the widely used Dublin Core metadata vocabulary but extended the vocabulary to make it more adequate for language resources (Bird & Simons 2001, 2003). The IMDI metadata format was developed roughly at the same time by the Max Planck Institute for Psycholinguistics in Nijmegen (Broeder et al. 2001). While not adopted as widely as OLAC, it provided a richer domain specific metadata vocabulary (Broeder & Wittenburg 2006). Later developments, such as the CMDI framework for metadata with its registries (Broeder et al. 2012) has built on this work on metadata and applied it to language resources beyond language documentation. More recently, a common tool for metadata entry, Lameta (Hatton et al. 2021), was developed. Continuing the efforts to provide technological tools to support documentary linguistics work.

The technical implementation of data and metadata formats has been constantly accompanied by a methodological discussion of what constitutes data and what metadata (e.g. Nathan & Austin 2004) and the discussion and definition of different data types (e.g. Himmelmann 2012). The dialogue between methodological discourse and infrastructure has clearly informed the technical implementations and experiences with the specific implementations has generated new methodological discussions.

1. https://www.delaman.org/about/ (accessed 17.07.2025)

4. Data quality and corpus creation

Data quality and principles of corpus creation were already central to Himmelmann (1998) and have remained central (see for example Evans & Sasse 2007; Johnson 2004; Woodbury 2007; Mosel 2018; Himmelmann 2018). The availability of high-quality collections of annotated audiovisual data through the publication in language archives has led to an increased engagement with these resources from a linguistic diversity perspective. The resources have allowed for the development of new approaches to comparative research. This utilization of collections from language archives in comparative typological research has produced new requirements for language resources and as a consequence the language documentation and linguistic diversity community has put more effort into standardization and comparability of the structural data in language documentation corpora. This resulted in corpora such as SCOPIC (Barth & Evans 2017), DoReCo (Paschen et al. 2020), and Multi-CAST (Schnell et al. 2021b), which are specifically designed and curated to facilitate corpus-based typology (Schnell et al. 2021a). Requirements for these types of resources have been incorporated into data quality guidelines and data quality assessments (Lange & Aznar 2022; Wamprechtshammer et al. 2022).

However, SCOPIC, Multi-CAST, and DoReCo have also shifted the emphasis from endangered languages to general comparative research questions in linguistics, while maintaining an emphasis on annotated high-quality audio-visual language corpora. This brings data types from documentary linguistics, their data quality criteria, as well as their research infrastructures to majority languages. Language archives are developing into institutions that hold audio-visual corpora of embodied situated language use from not only endangered language. In this new role, language archives become a research infrastructure for a rich type of annotated audio-visual data as an empirical base for general linguistics — a point already present in Himmelmann (2006).

Language archives have developed into institutions with a highly specialized competency in methodological and technical aspects of audio-visual language data and situate the practices, methodological concepts, and technical implementations in a wider context of research infrastructures.

5. In the research infrastructure landscape

Research infrastructures are "facilities that provide resources and services for the research communities to conduct research"[2] and

> can be mission-oriented and devoted to single disciplines, or multidisciplinary and open-ended in their use, and they can take a variety of technological shapes and forms including instruments and tools for discovery and experimentation, repositories of data and materials, and vessels for exploration.
>
> (Hallonsten 2020: 617)

Language archives developed as discipline (or sub-discipline) specific institutions and have increasingly integrated into a wider research infrastructure landscape.

A discipline-specific data ecosystem is not unique in the humanities. Other disciplines — such as archaeology (see for example Meghini et al. 2017; Hollander et al. 2017; Kreuzburg et al. 2023) — have developed considerable national and international infrastructures. The existence of a dedicated and professional discipline-specific infrastructure is not a matter of course and requires constant effort to maintain and further develop its parts. Documentary linguistics as a linguistic subdiscipline has developed and maintained a significant global infrastructure, which has contributed data and metadata standards as well as ethical and policy best practices to a wider landscape. In recent years, requirements from outside the discipline had a growing influence on this infrastructure. In particular, general principles for good data management and infrastructure, such as the FAIR (Findability, Accessibility, Interoperability, Reusability; Wilkinson et al. 2016) and the CARE principles (Collective Benefit, Authority to Control, Responsibility, Ethics; Carroll et al. 2020) had a global and cross-disciplinary influence on infrastructure discourse and development. However, there are also more regionally or theoretically restricted influences. In Europe, the CLARIN infrastructure (Bański & Hedeland 2022; Fišer & Witt 2022), which started out as a federation of trusted archive centers (Váradi et al. 2008), has advanced standardization across national and (sub-)disciplinary borders. These larger infrastructure networks allow for a homogenization of standards inside larger data domains and across disciplines. Language archives have participated early on in these larger research infrastructure projects (Drude et al. 2012) and the methodological discourse from documentary linguistics has influenced the wider infrastructure, in particular with regard to metadata and data access. The fact that Peter Wittenburg, the former

2. Regulation (EU) 2021/695 of the European Parliament and of the Council of 28 April 2021 establishing Horizon Europe http://data.europa.eu/eli/reg/2021/695/oj (accessed 21 September 2025)

head of The Language Archive in Nijmegen, is a co-author of the FAIR principles (Wilkinson et al. 2016) and of the FAIR Digital Objects (De Smedt et al. 2020) really highlights the deep integration of the language archive community in the wider landscape.

6. In a network of relations

Beyond being technical facilities that hold and provide high-quality digital research data and beyond being a research infrastructure that is part of a larger research infrastructure landscape, language archives as institutions exist in a network of relations. The archive, or the people who — through their work — constitute and perpetuate the archive, take part in a thick network of relationships. These relations connect the archive with the depositor/researcher, the speaker/signer, the relatives of the speaker/signer, the speech community and/or institutions representing the speech community. They also connect the archive with researchers working on the language or culture and may not have direct contact with the speech community. Furthermore, the archive as the institution preserving funded research output has a relation to the funders. As institutions preserving cultural heritage, the archives also may have a relation to the general public. These relationships — more than any technical component — define the archive and are crucial for the functioning of any language archive.

In particular, the relationship with the speakers and the speech community has been central for documentary linguistic work since the very beginning (Himmelmann 1998:172). Henke and Berez-Kroeker (2016) have identified their last and most recent phase of their history of language archiving as "Redefining archiving through participatory models (2010–)". The methodological discussion around ethics and the relation to speaker and speech community has started with Himmelmann (1998) who already raised "the question of how the communities can be actively involved in the design of a concrete documentation project from the very beginning" (Himmelmann 1998:188). While following discussions focused more on the relationship between the researcher and the speaker or the community (e.g. Austin 2010; Dobrin & Berson 2011), already Dwyer (2006) included language archives in *this web of relations* (Dwyer 2006:36). Language archives have addressed the relationship with speakers and communities (Holton 2012; Schwiertz 2012; Thieberger & Harris 2022), but maintaining and fostering relationships in an infrastructure and funding landscape that focuses more on digital objects and standards will remain a challenge for language archives. The increasing recognition of relationality in research practices (Wilson 2020; Tynan 2021), the complex nature of accountability and the role of community (Starblanket 2018;

Sullivan 2020) and in particular the role of relationality in language technologies (Meighan 2021; Chew et al. 2022) can guide and support language archives in their effort to become or remain relational and participatory institutions.

7. Conclusion and outlook

Modern language archives were formed as part of the development of documentary linguistics as a discipline with its own methodological discourse. They have found a place in a network of relations to researchers and speakers as well as in the research infrastructure landscape. The methodological requirements of language documentation, the accountability towards the speakers and the speech community as well as standards derived from the general research infrastructures shape the form and development of language archives. The archives have developed a unique set of services and competencies. They provide standardized rich domain-specific metadata, preserve annotated audio-visual language recordings and make these resources accessible, in an ethical and responsible way. Establishing and maintaining language archives as sustainable institutions has neither always nor everywhere been a simple and unquestioned endeavor. Still, as single archives as well as a global network, language archives have proven to be resilient institutions. Language archives manage relationships with many stakeholders and in particular with the speech communities. They preserve and provide valuable data for linguistic research and beyond.

The close connection of these research infrastructure institutions with the addressed research community as well as with other stakeholders will continue to enable a close and dynamic dialogue between conceptual and methodological discourse on the one hand and the further development of services and resources on the other. This deep integration into the theory and practices of documentary linguistics will ensure that language archives and the resources they preserve will remain relevant for scientific work on languages.

Himmelmann (1998) has introduced most major topics that influenced the directions into which the language archives developed over the following decades. Developments in linguistic and general research methods will continue to bring changes to how language archives develop. Documentary linguistics will remain the conceptual base for this kind of research infrastructure institutions.

Acknowledgements

I would like to thank Gabriele Schwiertz for initial discussions of the topic and her support to get the paper going, Birgit Hellwig for her extremely helpful comments on earlier versions of this paper and the whole editorial team for their patience.

References

Austin, Peter K. 2010. "Communities, Ethics and Rights in Language Documentation." *Language Documentation and Description* 7: 34–54.

Austin, Peter K. 2013. "Language Documentation and Meta-Documentation." In *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*, ed. by Mari C. Jones, and Sarah Ogilvie, 3–15. Cambridge: Cambridge University Press.

Bański, Piotr, and Hanna Hedeland. 2022. "Standards in CLARIN." In *CLARIN: The Infrastructure for Language Resources*, ed. by Darja Fišer, and Andreas Witt, 307–339. Berlin, Boston: De Gruyter Mouton.

Barth, Danielle, and Nicholas Evans (eds). 2017. *Social Cognition Parallax Interview Corpus (SCOPIC)*. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24739

Bechert, Johannes. 1990. "Universalienforschung und Ethnozentrismus." In *Proceedings of the Fourteenth International Congress of Linguists*, Vol. 3, ed. by Werner Bahner, Joachim Schildt, and Dieter Viehweger, 2350–2352. Berlin, Boston: De Gruyter Mouton.

Berez-Kroeker, Andrea L., and Ryan Henke. 2018. "Language Archiving." In *The Oxford Handbook of Endangered Languages*, ed. by Kenneth L. Rehg, and Lyle Campbell, 347–369. Oxford: Oxford University Press.

Bird, Steven, and Gary Simons. 2001. "The OLAC Metadata Set and Controlled Vocabularies." In *STAR '01: Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources*, Vol. 15, 7–18. Stroudsburg, PA: Association for Computational Linguistics.

Bird, Steven, and Gary Simons. 2003. "Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources." *Computers and the Humanities* 37 (4): 375–388.

Broeder, Daan, Freddy Offenga, Don Willems, and Peter Wittenburg. 2001. "The IMDI Metadata Set, Its Tools and Accessible Linguistic Databases." In *Proceedings of the IRCS Workshop on Linguistic Databases*, ed. by Steven Bird, Peter Buneman, and Mark Liberman, 11–13. Philadelphia: University of Pennsylvania.

Broeder, Daan, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. "CMDI: A Component Metadata Infrastructure." In *Proceedings of the Workshop Describing Language Resources with Metadata: Towards Flexibility and Interoperability in the Documentation of Language Resources (LREC 2012)*, ed. by Victoria Arranz, Daan Broeder, Bertrand Gaiffe, Maria Gavrilidou, Monica Monachini, and Thorsten Trippel, 1–4. Paris: European Language Resources Association.

Broeder, Daan, and Peter Wittenburg. 2006. "The IMDI Metadata Framework, Its Current Application and Future Direction." *International Journal of Metadata, Semantics and Ontologies* 1 (2): 119–132.

Broeder, Daan, Hennie Brugman, Albert Russel, and Peter Wittenburg. 2000. "A Browsable Corpus: Accessing Linguistic Resources the Easy Way." Presentation, Workshop Meta-Descriptions and Annotation Schemas for Multimodal/Multimedia Language Resources (LREC 2000), Athens, 29–30 May 2000. Available at https://www.mpi.nl/ISLE/documents/papers/broeder_paper.pdf

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. "The CARE Principles for Indigenous Data Governance." *Data Science Journal* 19 (November): 43.

Chew, Kari A. B., Lokosh (Joshua D. Hinson), and Juliet Morgan. 2022. "Centering Relationality in Online Indigenous Language Learning: Reflecting on the Creation and Use of Rosetta Stone Chickasaw." *Language Documentation and Conservation* 16: 228–258. https://hdl.handle.net/10125/74679

De Smedt, Koenraad, Dimitris Koureas, and Peter Wittenburg. 2020. "FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units." *Publications* 8 (2): 21.

Dobrin, Lise M., and Josh Berson. 2011. "Speakers and Language Documentation." In *The Cambridge Handbook of Endangered Languages*, ed. by Julia Sallabank, and Peter K. Austin, 187–211. Cambridge: Cambridge University Press.

Drude, Sebastian, Daan Broeder, Paul Trilsbeek, and Peter Wittenburg. 2012. "The Language Archive — a New Hub for Language Resources." In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 3264–3267. Paris: European Language Resources Association.

Dwyer, Arienne M. 2006. "Ethics and Practicalities of Cooperative Fieldwork and Analysis." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 31–66. Berlin, New York: De Gruyter Mouton.

Evans, Nick, and Hans-Jürgen Sasse. 2007. "Searching for Meaning in the Library of Babel: Field Semantics and Problems of Digital Archiving." *Language Documentation and Description* 4: 58–99.

Fišer, Darja, and Andreas Witt (eds.). 2022. *CLARIN: The Infrastructure for Language Resources*. Berlin, Boston: De Gruyter Mouton.

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. 1992. "Endangered Languages." *Language* 68 (1): 1–42.

Hallonsten, Olof. 2020. "Research Infrastructures in Europe: The Hype and the Field." *European Review* 28 (4): 617–635.

Hatton, John, Gary Holton, Mandana Seyfeddinipur, and Nick Thieberger. 2021. *Lameta*. [Computer program]. https://github.com/onset/laMETA/releases

Henke, Ryan, and Andrea L. Berez-Kroeker. 2016. "A Brief History of Archiving in Language Documentation, with an Annotated Bibliography." *Language Documentation and Conservation* 10: 411–457. http://hdl.handle.net/10125/24714

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–196.

Himmelmann, Nikolaus P. 2006. "Language Documentation: What Is It and What Is It Good For?" In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin, New York: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2012. "Linguistic Data Types and the Interface between Language Documentation and Description." *Language Documentation and Conservation* 6: 187–207. http://hdl.handle.net/10125/4503

Himmelmann, Nikolaus P. 2018. "Meeting the Transcription Challenge." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 33–40. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24806

Hollander, H. S., Nicola Aloia, Ceri Binding, Sebastian Cuy, Martin Doerr, Bruno Fanini, Achille Felicetti, Johan Fihn, Dimitris Gavrilis, Guntram Geser, Carlo Meghini, Franco Niccolucci, Federico Nurra, Christos Papatheodorou, Julian Richards, Paola Ronzino, Roberto Scopigno, Maria Theodoridou, Maria Theodoridou, Douglas Tudhope, Andreas Vlachidis, and Holly Wright. 2017. "Enabling European Archaeological Research: The ARIADNE E-Infrastructure." *Internet Archaeology* 17 (43): 43.

Holton, Gary. 2012. "Language Archives: They're Not Just for Linguists Any More." In *Potentials of Language Documentation: Methods, Analyses, and Utilization*, ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek, 111–117. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/4523

Johnson, Heidi. 2004. "Language Documentation and Archiving, or How to Build a Better Corpus." *Language Documentation and Description* 2: 140–153.

Kreuzburg, Lea, Juliane Watson, Fabian Riebschläger, and Alexander Dorn. 2023. "Zentral und wartbar sollst Du sein! Der FDM-Bereich am DAI. Die Erstellung eines Informationshubs für das Forschungsdatenmanagement. [You Should Be Centralized and Maintainable! The FDM Department at the DAI. Creating an Information Hub for Research Data Management]" *Forum for Digital Archaeology and Infrastructure:* 1–57 (§).

Lange, Herbert, and Jocelyn Aznar. 2022. "RefCo and Its Checker: Improving Language Documentation Corpora's Reusability Through a Semi-Automatic Review Process." In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, 2721–2729. Paris: European Language Resources Association.

Lehmann, Christian. 2001. "Language Documentation: A Program." In *Aspects of Typology and Universals*, ed. by Walter Bisang, 83–98. Berlin: Akademie Verlag.

McGregor, William. 2008. "History of Fieldwork on Kimberley Languages." In *Encountering Aboriginal Languages: Studies in the History of Australian Linguistics*, ed. by William McGregor, 1–34. Canberra: Pacific Linguistics.

Meghini, Carlo, Roberto Scopigno, Julian Richards, Holly Wright, Guntram Geser, Sebastian Cuy, Johan Fihn, Bruno Fanini, Hella Hollander, Franco Niccolucci, Achille Felicetti, Paola Ronzino, Federico Nurra, Christos Papatheodorou, Dimitris Gavrilis, Maria Theodoridou, Martin Doerr, Douglas Tudhope, Ceri Binding, and Andreas Vlachidis. 2017. "ARIADNE: A Research Infrastructure for Archaeology." *Journal on Computing Cultural Heritage* 10 (3): 18.

Meighan, Paul J. 2021. "Decolonizing the Digital Landscape: The Role of Technology in Indigenous Language Revitalization." *AlterNative: An International Journal of Indigenous Peoples* 17 (3): 397–405.

Mosel, Ulrike. 2018. "Corpus Compilation and Exploitation in Language Documentation Projects." In *The Oxford Handbook of Endangered Languages*, ed. by Kenneth L. Rehg, and Lyle Campbell, 248–270. Oxford: Oxford University Press.

Nathan, David, and Peter K. Austin. 2004. "Reconceiving Metadata: Language Documentation through Thick and Thin." *Language Documentation and Description* 2: 179–188.

Paschen, Ludger, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. "Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo)." In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2657–2666. Paris: European Language Resources Association.

Rice, Keren, and Nicholas Thieberger. 2018. "Tools and Technology for Language Documentation and Revitalization." In *The Oxford Handbook of Endangered Languages*, ed. by Kenneth L. Rehg, and Lyle Campbell, 225–247. Oxford: Oxford University Press.

Schnell, Stefan, Geoffrey Haig, and Frank Seifart (eds). 2021a. *Doing Corpus-Based Typology With Spoken Language Corpora: State of the Art*. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/74663

Schnell, Stefan, Geoffrey Haig, and Frank Seifart. 2021b. "The Role of Language Documentation in Corpus-Based Typology." In *Doing Corpus-Based Typology with Spoken Language Data: State of the Art*, ed. by Geoffrey Haig, Stefan Schnell, and Frank Seifart (eds.), 1–28. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/74656

Schwiertz, Gabriele. 2012. "Online Presentation and Accessibility of Endangered Languages Data: The General Portal to the DoBeS Archive." In *Potentials of Language Documentation: Methods, Analyses, and Utilization*, ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek, 126–128. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/4526

Seifart, Frank, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. "Language Documentation Twenty-Five Years On." *Language* 94 (4): e324–e345.

Seifart, Frank, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (eds). 2012. *Potentials of Language Documentation: Methods, Analyses, and Utilization*. Honolulu: University of Hawai'i Press. https://hdl.handle.net/10125/23013

Sloetjes, Han, and Peter Wittenburg. 2008. "Annotation by Category: ELAN and ISO DCR." In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, 816–820. Paris: European Language Resources Association.

Starblanket, Gina. 2018. "Complex Accountabilities: Deconstructing 'the Community' and Engaging Indigenous Feminist Research Methods." *American Indian Culture and Research Journal* 42 (4): 1–20.

Sullivan, Corrinne Tayce. 2020. "Who Holds the Key? Negotiating Gatekeepers, Community Politics, and the 'Right' to Research in Indigenous Spaces." *Geographical Research* 58 (4): 344–354.

Sullivant, Ryan. 2020. "Archival Description for Language Documentation Collections." *Language Documentation and Conservation* 14: 520–578. hdl.handle.net/10125/24949

Thieberger, Nick. 2016. "Documentary Linguistics: Methodological Challenges and Innovatory Responses." *Applied Linguistics* 37 (1): 88–99.

Thieberger, Nick, and Amanda Harris. 2022. "When Your Data Is My Grandparents Singing. Digitisation and Access for Cultural Records, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)." *Data Science Journal* 21 (1): 9.

Trilsbeek, Paul, and Peter Wittenburg. 2006. "Archiving Challenges." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 311–336. Berlin, New York: De Gruyter Mouton.

Tynan, Lauren. 2021. "What Is Relationality? Indigenous Knowledges, Practices and Responsibilities with Kin." *Cultural Geographies* 28 (4): 597–610.

Váradi, Tamás, Steven Krauwer, Peter Wittenburg, Martin Wynne, and Kimmo Koskenniemi. 2008. "CLARIN: Common Language Resources and Technology Infrastructure." In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, 1244–1248. Paris: European Language Resources Association.

Wamprechtshammer, Anna, Jocelyn Aznar, Elena Arestau, Hanna Hedeland, Amy Isard, Ilya Khait, Herbert Lange, Nicole Majka, and Felix Rau. 2022. "QUEST: Guidelines and Specifications for the Assessment of Audiovisual, Annotated Language Data." *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology* 8: 1–85.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018.

Wilson, Shawn. 2020. *Research Is Ceremony: Indigenous Research Methods*. Nova Scotia: Fernwood Publishing.

Wittenburg, Peter, Ulrike Mosel, and Arienne M. Dwyer. 2002. "Methods of Language Documentation in the DOBES Project." *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, ed. by Manuel González Rodríguez, and Carmen Paz Suarez Araujo, 36–42. Paris: European Language Resources Association.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. "ELAN: A Professional Framework for Multimodality Research." In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, ed. by Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, 1556–1559. Paris: European Language Resources Association.

doi Woodbury, Anthony C. 2003. "Defining Documentary Linguistics" *Language Documentation and Description* 1: 35–51.

doi Woodbury, Anthony C. 2007. "On Thick Translation in Linguistic Documentation." *Language Documentation and Description* 4: 120–135.

CHAPTER 7

Science in public

The representation of *language documentation* in science communication

Tobias-Alexander Herrmann, Kai Kaspar
& Stephan Packard
University of Cologne

How do public media represent language documentation? We offer a first pilot study informed by the Public Understanding of Science framework and recent media history. We employ a mixed-methods approach on a corpus of selected media coverage, using Topic Modelling to identify and quantify recurring topics, and corpus-assisted discourse analysis in the form of KWIC and collocations for qualitative analysis. We show that only a good one in two mentions of language documentation in media coverage corresponds to the way it is accepted in academia, whereas outlets tend to highlight certain characteristics of the discipline, overemphasise certain facets, and sometimes ignore relevant strands. In addition, there are competing meanings in discourse which belong to unrelated fields.

Keywords: language documentation, science communication, public understanding of science, media representation, media discourse, topic modelling

1. Introduction

The societal significance of language documentation extends far beyond academia, for instance in terms of language endangerment and language diversity. Here, public engagement and awareness are crucial for ensuring the long-term success of documentation efforts, meaning the academic discipline must rely on public support and awareness for funding, collaboration, and impact. In this regard, understanding public media narratives is critical to meet the goals of fostering public trust, securing funding, and ensuring broad-based support for linguistic research. However, it remains unclear how language documentation is actually portrayed in public media, particularly in science communication, and which

https://doi.org/10.1075/slcs.240.07her
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

aspects are emphasised. We offer a first pilot study that approaches this research gap, putting forward the following research question: How is language documentation represented in news media? While linguistic and corpus analyses allow for substantial insights into the status quo of language documentation in science communication, many of the associated questions demand an interdisciplinary view that can begin to connect the detailed description of the surface message with the conditions of its production and the likely horizon of its interpretation by a heterogeneous audience. This contribution tackles this challenge by combining a detailed linguistic analysis with concepts relating to the shifting conditions of mass communication from critical media studies as well as standards from media psychology.

The contribution is structured as follows: We begin by presenting the goals and concepts of language documentation in science communication in Section 1.1, and then move on to sketch the multi-theoretical Public Understanding of Science (PUS) framework in Section 1.2 in which we ground our analysis. This enables us to examine how the scientific field of language documentation is communicated in public contexts, including the media. In Section 1.3 we disentangle the cover term *media* and discuss recent shifts in public science communication to determine which types of outlets are relevant for the present study, before zooming in on our pilot study in Section 1.4. Turning to Section 2.1, we outline the data collection and preprocessing steps, explain the study's methodology in Section 2.2, and describe how we conducted the analyses to address our research question in Sections 2.3 and 2.4, respectively. The results are presented in Section 3 and discussed in Section 4, before we wrap up the contribution with concluding remarks in Section 5.

1.1 Language documentation in science communication: Goals and concepts

Language documentation and documentary linguistics can be defined

> as a field of linguistic inquiry and practice in its own right which is primarily concerned with the compilation and preservation of linguistic primary data and interfaces between primary data and various types of analyses based on these data. (Himmelmann 2006: 1)

A language documentation consists of a corpus of recordings of observable linguistic behaviour and meta-linguistic knowledge as well as metadata and annotations (Himmelmann 2006: 14). The process of collecting primary data is considered language documentation, and is distinguished from providing a descriptive analysis of these data, which is language description (Himmelmann 1998).

The relevance of the academic discipline of language documentation is underscored by factors such as "language endangerment, the economy of research resources, and accountability" (Himmelmann 2006: 5). Language corpora and repositories aid in language maintenance while also strengthening the empirical basis of various linguistic and interdisciplinary research projects and "related disciplines which heavily draw on data of little-known speech communities (e.g. linguistic typology, cognitive anthropology, etc.) in that they significantly improve accountability (verifiability) and economizing research resources" (Himmelmann 2006: 1).

In the context of public trust in science, research outcomes have to be meaningful to a wider audience, which is why the methodological foundations and objectives of the discipline must be accessible and well understood (cf. Durant e al. 1989). In addition, Linn (2018: 66) highlights the need to raise "public awareness of language endangerment" and stimulate "global attitudinal changes in support of language diversity". She emphasises that public awareness of the significance of language documentation is crucial for the success of documentation efforts and, consequently, for preventing the widespread loss of languages. Public support for science, in turn, is closely linked to funding opportunities because research areas with greater public relevance and use tend to attract more funding (e.g. Yin et al. 2021, 2022; Caballe & Bardelli 2021). Fortunately, dedicated funding programmes for endangered languages have gained visibility in policy decisions and international organisations such as UNESCO and the European Union (Czaykowska-Higgins 2018: 115). However, there remains a gap between available resources and the actual need, with Holton and Seyfeddinipur (2018: 106) arguing that the number of funded projects is too small, which is why "linguists will need to look elsewhere to fund language documentation work in the future". To bridge this gap, researchers must increasingly engage in broader outreach efforts.

As the public discourse plays a central role in constructing public belief and attitudes and its relationship to social change (Happer & Philo 2013) as well as shaping how science is perceived, supported, and funded — particularly as media helps to disseminate research results to the general public (e.g. Eagleman 2013) —, understanding the media representation and public perception of language documentation is key to addressing broader questions of visibility and legitimation.

When it comes to language documentation, there is a notable absence of studies investigating how the discipline is represented in public discourse. All of these questions have to be considered in the context of a rapidly changing public sphere (Fraser 2007; Imhof 2011; Habermas 2022). As television had previously begun to outreach print journalism as the main source of science communication, the rise of 24-hours news media in the 1990s redefined the main usage parameters of asymmetric, few-to-many media productions. The subsequent shift to web for-

mats and ultimately social media platforms might constitute an even more radical transformation, as many-to-many constellations allow for — but often also demand — direct communication between scientists and the public at large, either redefining or possibly marginalising the role of science communicators. The basic tenets of science communication and especially the question of a precarious balance between simplification, empowerment, and misdirection (Hilgartner 1990; Bonfadelli 2017) have to be reconsidered in this light.

Hence, several points of interest relevant for researchers from documentary linguistics are unclear. Do the media represent the difference between language documentation and language description as is accepted in academia? How is the relevance of the discipline portrayed and justified in public discourse? What are key aspects in connection with media coverage on documentary linguistics? Do media outlets actually pick up on scientific results obtained in documentary linguistics and their impact? Answering all of these questions is of course beyond the scope of the present contribution. However, this pilot study approaches the overall research gap, putting forward the research question of how language documentation is represented in the media.

1.2 Public Understanding of Science

The term Public Understanding of Science (PUS) is best described as a broad framework that brings together diverse theoretical strands, including cognitive, sociological, and communication models, to address how science is represented and interpreted in public contexts (Gregory & Miller 1998; Brossard & Lewenstein 2010; Yearley 1994). Our basic assumption is that scientific results are not simply obtained and then shared and communicated in the media in a neutral fashion, but that science is discursively constructed. This applies to language documentation as well, meaning it is produced, communicated, represented, and interpreted in the media, ultimately contributing to shaping public understanding of the discipline (Farr 1993). In this context, Millar and Wynne (1988) as well as Wynne (1992) argue that PUS should address not just knowledge, but must go beyond knowledge transmission to include how science is organised, constructed, controlled and interpreted in public discourse, particularly through media representation. As a consequence, the PUS framework sheds light on key questions that need to be addressed when examining the public discourse of language documentation. These include which topics dominate, which actors are present and whose voices are amplified as well as what frames, narratives, and interpretive patterns emerge. Hence, addressing our research question of how language documentation is represented in the media, we ground our analysis in the theoretical core of PUS.

1.3 "The" media: Shifting concepts of a public sphere

Prior research has investigated how academic disciplines are represented in public media (see e.g. the meta-analysis of studies on the media's coverage of science by Schäfer 2012), highlighting the value of examining how academic fields are constructed in public discourse, although hitherto none has focused on works specifically on language documentation. Among the central issues in the literature are the misrepresentation of academia in public media and the gaps between them. For instance, Sumner et al. (2014) found that news media strongly exaggerated scientific results, showing that this was associated with the way the evidence was presented in press releases. Peters (2013) pointed to the gap between the arenas of internal scientific and public communication, and Mach et al. (2021) found moderate scientific quality in COVID-19 media reports. Meanwhile, Yu et al. (2025) demonstrated that public sources to some extent oversimplified and misreported original research results, providing conclusions that were not backed by the actual scientific evidence. Likewise, Buzova et al. (2024) compared how a specific economic concept is framed in academic literature versus news media, revealing significant differences.

Another central issue is a bias in public representation, with media sources selecting certain disciplines and topics. For instance, Ampollini and Bucchi (2019) analysed how public media coverage of research integrity mirrors academic debates, showing that integrity and misconduct alongside public trust in science are salient themes in the news. Jonker and Vanlee (2024), meanwhile, examined public media mentions and policy citations of academics across disciplines, revealing differences between disciplines in terms of visibility, and a representational skew whereby publications frequently cited in academia were rarely featured in traditional media.

A further aspect concerns how public science communication sometimes focuses on individual academics to provide authoritative insights. On the one hand, scientists may actively steer their role in public communication, including constructing the public image of science (Horst 2013). On the other hand, scientists' portrayal in public media may affect the representation of an academic discipline, e.g. in terms of "human" or "objective" faces of science (Mitchell & McKinnon 2019) or in terms of gender and sexualisation (Attenborough 2013).

So far, we have demonstrated that *the media* take on a central role in managing and producing public discourse. However, this term obscures the fact that we are not dealing with a single, monolithic entity, but are referring to a collection of different organisations and individuals involved in the dissemination of information, which is subject to change over time. As public science communication is traditionally tied to the concept of a common, shared public sphere, issues of pop-

ularisation are connected to the challenges of interdiscursivity (Hilgartner 1990; Jäger 2012), i.e. the negotiation of communicative contact between one or several highly specialised, in this case academic, discursive arrangements, and a heterogeneous public brought together by mass media. Science communicators have been charged with the task of balancing accuracy, intelligibility, and mass media reach as they work with experts to communicate to laypeople.

Habermas (2022) famously extolled the virtues of an actively and passively accessible public sphere but decried the influence of television as distorting rational discourse with appeals to popularity and mass communication. Pierre Bourdieu's unique sociological interventions into televised mass media, *Sur la télévision* and *Le champ journalistique*, were first presented as monologues in an unprecedented televised format. They have, however, been almost exclusively cited and discussed in their subsequent printed form (1996). These works offer a more concrete vision of the problematic position of expert communication through mass media, as they heavily criticised television as a whole by addressing both general and contemporarily acute challenges for science communication through journalism. Bourdieu's criticism structurally depends on the duplication of the academic or scholarly field by the field of journalism. By stepping in between specialised and popular discourses, this field distorts evaluations, controversies, and hierarchies. A discursively advanced understanding by an audience of limited extent is replaced with engagement by a much larger audience that may ultimately fail to understand the issues at hand altogether. Linking this perspective to the acknowledged difference between language documentation and language description in academia, it can be predicted that systematically separate branches of linguistic practice might fare differently as they undergo transportation into the separate field of journalism. At least, it is justified to ask how well the shift from academic representation to the mediating functions of journalism works out.

That translation has been subjected to at least two major shifts in recent media history, within the scope of the material discussed in this study. Public media and journalism might be divided chronologically into three connected and blurred, yet typologically distinct, conditions of communication that might be somewhat simplistically labelled as 24-hour television, networked digitisation, and post-digitality. First, the introduction of 24-hour news cycles, beginning in the early 1980s and reshaping Western international media consumption no later than the early 1990s (cf. Gilboa 2005), emphasised the role of television for the informative and critical functions of public discourse; foregrounded the previously neglected entertainment as the main dimension of audience gratification in news media and journalism; and resulted in a redifferentiation of printed journalism that intensified the latter's critical and reflective role. A threshold, usually dated roughly to

CNN's coverage of the Gulf War in 1991, marks a shift in media sources. Consequently, journalism before and after 1991 are very difficult to compare.

Similarly, the sharp distinction between online sources and otherwise networked material advises us against comparing formats from television or traditional newspapers to online textual sources. For this study, we focus on the latter. In the 1990s, the shift towards 24-hour television news cycles is coupled with the concurrent shift towards yet another emerging infrastructure for additional, still different public spheres that arise on the internet (cf. Jarren et al. 2000). These new public spheres departed from the asymmetry of traditional mass media in print, which was reinforced in television. Instead, the rise of internet platforms, which would eventually include recent social media formats, created a sharper contrast between journalistic science communication and laypeople's discussion of science, with the latter more regularly engaging with original sources in scientific discourse or even recruiting scientists as active agents in public debate without journalistic interference (cf. Luzón & Pérez-Llantada 2019). At the same time, the quantitative self-representation and self-observation of this new discourse skyrocketed due to the new Big Data technologies integrated into, and commercially exploited by, new platforms (cf. boyd & Crawford 2012). Following the argument in José van Dijck's seminal paper on the ensuing shift in science communication (2003), we might speculate that this new distinction between traditional science journalism and newly forged mixed public spheres subverted the previous distinction between experts and laypeople to an extent that it also suspended C. P. Snow's (1993) idea of the "two cultures" dominating different areas of scholarly inquiry. As scientists and their papers face direct engagement in new networked publics, those publics are faced with both qualitative and quantitative data collection and representation in real time feeding back into the algorithmic curation of their ongoing communication. This makes it likely that any direct comparison between journalism's treatment of language documentation between 1991 and 2000 to any earlier or later data would be obstructed by the constellation of completely different formats defining the contexts of each science communication effort.

2000 offers itself as a rough threshold for this reason: If the revolution of journalism through 24-hour television was fully realised by 1991, the following age of early digitisation is often said to have lasted very briefly, undergoing its own massive shift towards "post-digitality" around the millennium. For instance, Kim Cascone (2000) demonstrates that the proliferation of digital tools led to a democratisation of media production and engagement. A new media culture described by Pepperell and Punt (2000) as a confusion of actual technology, users' often divergent imagination of that technology, and their desires for that technology, rendered the distinction between "analogue" media and "new digital" media obsolete.

All told, the take-away message is that we consider *the media* as a complex ecosystem of actors and influences that has seen drastic changes recently. Although we use the term throughout the paper, we acknowledge the diversity of both media and audiences. For analytical purposes, however, we operationalise the public discourse with news articles and press releases published mainly online during the past 25 years, mirroring the time span of digitisation and post-digitality. Being aware that this selection of news coverage may limit the study's validity, we believe it is the right step for a pilot study that is about to enter a new avenue of research.

1.4 The current study

Having laid out the PUS framework and devoted theoretical considerations to the complexity of *the media*, this section narrows down the focus of the present pilot study.

To recap, our research question is how the academic discipline of language documentation is represented in the media. Rooted in the PUS framework, relevant issues that must be examined are: Which topics dominate, which actors are present, and what frames and narratives emerge in the public discourse of language documentation? Hence, to address this research question, we draw on a mixed-methods approach, conducting Topic Modelling (TM) and a corpus-assisted discourse analysis, specifically looking at keywords in contexts and collocations. TM allows to systematically identify and quantify the main themes and topics that structure media coverage of a scientific field, such as language documentation. For instance, Jacobs and Tschötschel (2019) show that TM can enhance discourse analysis by scaling up the identification of dominant themes and patterns, without violating the epistemological premises of discourse theory. They argue that this approach is well-suited for systematically investigating how discourses are constructed in large media corpora. Hence, TM meets the PUS interest, in which topics are emphasised and how science is positioned in public discourse. Brookes and McEnery (2019) consider TM useful for mapping thematic structures, however, they highlight the need to pair TM with additional discourse analysis to fully capture interpretive patterns. This kind of approach again echoes PUS's emphasis on meaning-making and framing. Drawing on such a mixed-methods approach, Aranda et al. (2021) demonstrate that TM can be integrated with critical discourse analysis to capture both the breadth (what topics are present) and depth (how topics are framed and legitimated) of public debates about science, directly operationalising PUS concerns. An analysis of keywords in context and collocations provides the opportunity to go beyond surface-level themes to examine how key terms — in our case *language documentation* — are framed, what associations and evaluative patterns are present, and which actors

or perspectives are foregrounded. This aligns with PUS's focus on representation, framing, and interpretation. Bednarek (2024) reviewed the use of TM in corpus-based discourse analysis, highlighting its value for exploring which topics and frames dominate media representations of scientific issues.

All told, the previous literature demonstrates that TM and a corpus-assisted discourse analysis of keywords in context and collocations are a suitable means to address our research question to explore framing, directly operationalising the PUS framework's focus. In light of the notable absence of previous studies devoted to the representation of language documentation in the media as well as the study's methodology, our analysis does not test specific hypotheses but is explicitly exploratory. We move on to describe the methods as well as the data for the present pilot study in greater detail in the next section.

2. Materials and methods

2.1 Data collection and preprocessing

Media coverage on language documentation was exemplified by compiling a corpus of media articles which were retrieved via the Nexis Uni database (LexisNexis, advance.lexis.com). Nexis Uni contains data from over 17,000 news, corporate and legal sources. The news articles are from several thousand newspapers, journals, magazines and news wires from all over the world. In the database, using the term *language documentation* a simple search for English language news articles was performed which — in line with our theoretical considerations on the era of digitisation and post-digitality — were published between January 1, 1999 and December 31, 2024, thus spanning a period of 25 years. Of the 978 results, 118 articles were excluded because they were either not in English, did not contain the search term but were separated by a comma ("language, documentation"), contained the search term in a link on the website unrelated to the original source, or were not a news article (e.g. a scientific paper). The majority of articles in the corpus were published by news agencies and commercial news wires (e.g. *Targeted News Service, States News Service, Business Wire, PR Newswire*). While some articles come from internationally renowned traditional newspapers and magazines such as *The New York Times* and *The Christian Science Monitor* (U.S.), *The Times* (U.K.), *The Sunday Mail* (Australia) as well as *The Hindu* (India), they represent a minor share compared to the predominance of agency material.

The matching articles were downloaded as HTML files and subsequently processed using R (R Core Team 2024, 4.4.2). A custom web scraping script, developed with the *rvest* package (Wickham 2024), was used to extract relevant

metadata, including the title, publication outlet, publication date, and full-text content of each article. The script identified key HTML elements within the downloaded files and extracted structured information accordingly. In this fashion, a corpus comprising 860 articles was compiled.

Prior to the analysis, the collected articles underwent a preprocessing phase to clean the text, using several steps. Applying tokenisation first, the text was converted into individual words using the unnest_tokens() function from *tidytext* (Silge & Robinson 2016). Second, using the same package, standard English stopwords were removed. In addition, to prevent the model from focusing on the search term, which appeared in all documents by design, the words *language* and *documentation* were excluded prior to modelling as well, ensuring that topic discovery was not biased by the initial selection criterion. All URLs were removed from the articles' body text using regular expressions. Third, words were lemmatised using the *textstem* package (Rinker 2018) to reduce words to their base forms. As a result, after stopword removal and preprocessing a corpus comprising 444,191 words was compiled and used for topic modelling.

2.2 Topic modelling

TM is a foundational technique in text mining and natural language processing. It is a class of unsupervised machine learning methods developed to uncover latent thematic structures in large collections of text documents. These probabilistic models take as input a corpus of documents and return a set of topics — defined as clusters of related words — that represent the semantic content of the corpus in a compact and interpretable form (Churchill & Singh 2022; Cho 2019; Abdelrazek et al. 2023). As the approach allows to uncover hidden thematic structures in large document collections, it is suitable to examine the structure and organisation of the present study's data, representing the media discourse on language documentation.

The most widely used algorithm in TM is Latent Dirichlet Allocation (LDA), which assumes that each document is a mixture of multiple topics in different proportions (Blei 2012: 78). For instance, "in a document that is 10% about 'A' and 90% about 'B', there would probably be about 9 times more 'B' words than 'A' words" (Cho 2019: 115). The topics themselves are not defined in advance but are inferred from word co-occurrence patterns in the data. Despite the method's limitations (see e.g. Vayansky & Kumar 2020), TM provides an interpretable, latent representation of documents by associating them with distributions over the discovered topics. These topics in turn capture recurring themes, which may also vary in prominence across documents and evolve over time (Blei 2012; Abdelrazek et al. 2023). Since the modelling is unsupervised, it does not require manual annotation

or predefined categories, making it particularly useful for exploratory analysis of large, unstructured text corpora as is the case in the present study.

2.3 Implementation and procedure

All analyses were performed using R (R Core Team 2024, 4.4.2). The preprocessed text was transformed into a document-term matrix using the cast_dtm() function from the *topicmodels* package (Grün & Hornik 2011, 2024). This matrix represents the frequency of terms across different articles. The LDA model was then applied to uncover latent topics within the corpus.

Given the absence of strong theoretical priors due to a lack of previous research, the optimal number of topics was determined using the log-likelihood criterion. A series of LDA models with topic numbers ranging from 2 to 10 was computed, and the log-likelihood values were compared. Relative gain, calculated as the change in log-likelihood divided by the absolute log-likelihood of the previous model, was used to assess improvements. The most substantial improvements occurred when increasing the number of topics from 2 to 3 (relative gain of 2.18%), to 4 (relative gain of 1.55%), to 5 (relative gain of 1.07%), and to 6 (relative gain of 1.12%) with diminishing returns thereafter. Since the relative improvement in model fit decreased considerably after six topics, additional topics likely add only marginal explanatory power. Against the goal of achieving a more interpretable and generalisable solution, the number of topics was set to six.

The model was trained using the LDA() function from the *topicmodels* package, with a random seed of 1234 for reproducibility. The most relevant words for each topic were extracted by computing the term probabilities, and the top words for each topic were visualised using bar charts, where terms were ranked by their probability scores.

To explore how the prominence of topics evolved over time within the corpus, the posterior probabilities of topic assignment for each document were extracted from the fitted LDA model. These probabilities — referred to as gamma values — indicate the proportion of each topic present in a given document. The tidy() function (from the *tidytext* package) was used to extract the gamma matrix, which represents the topic distribution per document.

The average topic probability per month was computed across all documents to reveal how topics are distributed across time. To visualise these temporal dynamics, the average topic proportions were plotted over time.

All plots were generated using the *ggplot2* package (Wickham 2016).

2.4 KWIC and collocation analysis

To get a more detailed understanding of the linguistic characteristics and discourse patterns associated with specific topics, we paired TM with a corpus-assisted discourse analysis. To this end, focused subcorpora were created based on topic dominance using the output of the LDA model, which includes the topic proportion (gamma) for each document. For this purpose, documents were selected if they exhibited a gamma value of 0.9 or higher for a given topic, indicating strong topic alignment.

Subsequently, a more fine-grained textual analysis was performed on the five subcorpora using the *quanteda* package (Benoit et al. 2018, 4.3.1), conducting two complementary analyses to explore thematic and collocational structures. First, identifying instances of *language documentation* using a Keyword-in-Context (KWIC) analysis allowed for a contextual examination of how the compound was embedded in discourse. Second, extracting collocations using the textstat_collocations() function containing *language documentation* and consisting of multiword expressions occurring at least five times were isolated to explore recurring lexical patterns associated with the topic. These sequences co-occur more frequently than would be expected by chance. Based on these metrics, select articles were then examined qualitatively.

3. Results

This section presents the results of our analyses. We begin by reporting top terms, keywords in context and collocations, before examining topic trends over time.

3.1 Top terms and topics

To recall, TM is an exploratory method that helps to structure a large corpus without any prior assumptions, providing an overview of the overarching themes and their linguistic characterisation by means of keywords. This section presents the identified relevant keywords for each of the six topics. The top terms were extracted by computing the term probabilities. They are presented in the corresponding order grouped by topic in Figure 1. The KWIC and collocational analyses of each topic follow suit to qualitatively get hold of the data.

Topic 1. The top terms are *university, research, application, professor, include, code, science, patent, program, development, student, study, support, award* and *project*. These keywords point to *language documentation* occurring in a research-oriented context with a focus on applied linguistics. The most frequent colloca-

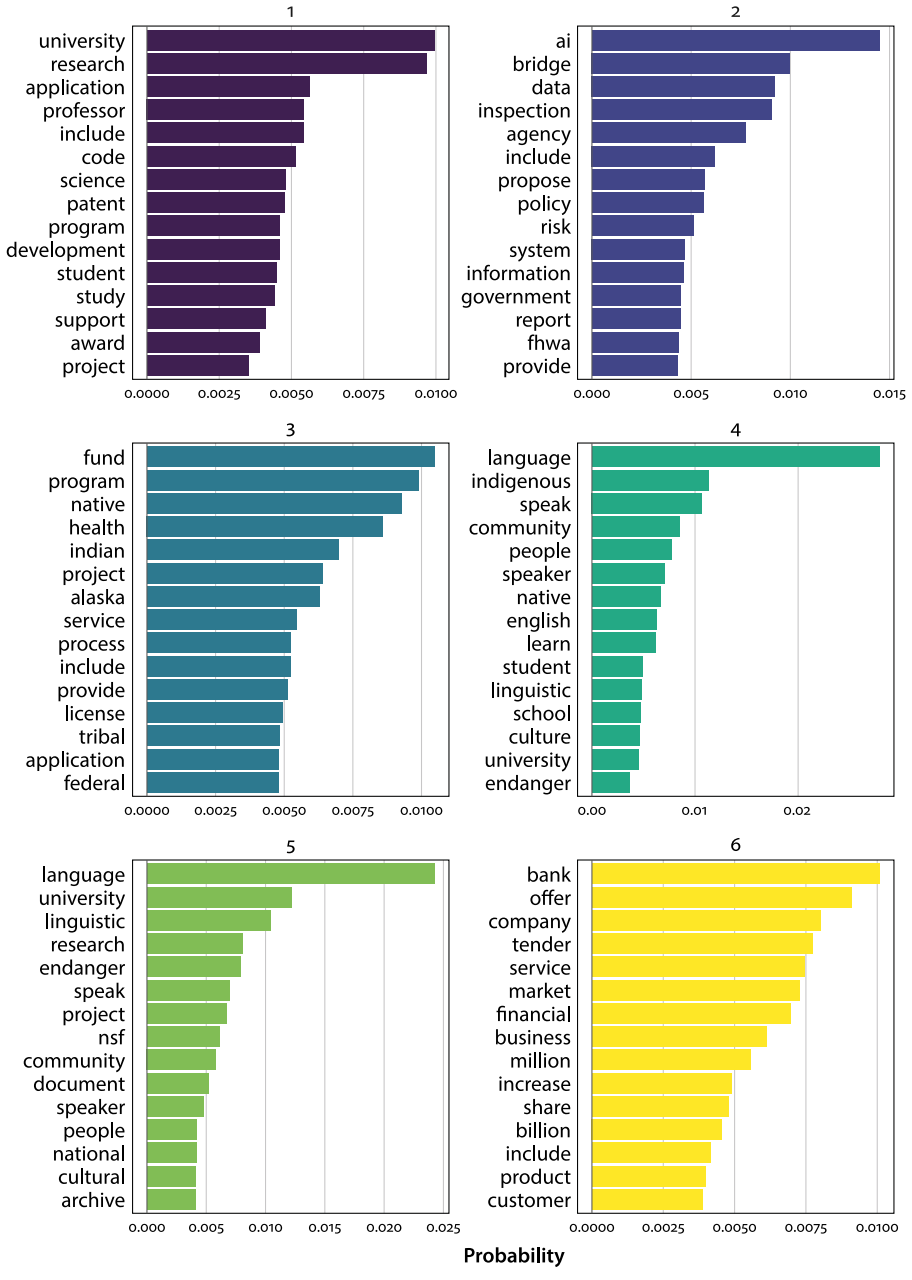


Figure 1. Top terms grouped by topics

tions from the topic's subcorpus (although statistically not significant: $\lambda = 0.93$, $z = 0.30$) is the bigram *language documentation and*, most often complemented by *revitalisation*. To illustrate the topic's contexts, compare the following examples. For instance, there is news coverage on several linguists, either receiving awards or taking on new roles at the university. Most outlets describe their expertise, mentioning *language documentation* in this context.

(1) She has been involved in the collaborative language documentation and revitalization projects of several Indigenous languages.[1]

(2) Her research investigates the phonology of Native American languages and Indigenous language documentation and revitalization.[2]

(3) Her research questions examine word formation, word meaning, the characterization of morphological and other categories, and general language documentation and revitalization.[3]

Another frequent collocation is the trigram *natural language documentation*. Compare the following excerpts exemplifying the context.

(4) The autoregressive language model may be trained on corpora of source code and natural language documentation on the subject of computer programming.[4]

(5) Miami, Florida has been assigned a patent […]. Systems and methods for automatically generating natural language documentation for source code files are provided.[5]

The examples make it clear that two types of meanings are included in the topic. On the one hand, language documentation refers to the research field. Here, it is presented in the context of individuals who are experts, and it is thus clearly framed as an academic discipline. On the other hand, the phrase may also describe reference guides and software architecture. Hence, the topic reveals a semantic split in the use of language documentation: it denotes both scholarly

1. *UWM Alumni Association announces 2017 Alumni Award winners* July 26, 2017. *Targeted News Service*. Accessed via Nexis Uni.

2. *Vice president for research and creative activity named* Aug. 19, 2021. *US Fed News*. Accessed via Nexis Uni.

3. *Faculty receive 2023–24 WARF named professorships, Kellett fellowships, and Romnes awards* June 13, 2023. *States News Service*. Accessed via Nexis Uni.

4. *US patent issued to X Development on Oct. 25 for "Conditioning autoregressive language model to improve code migration"* Oct. 26, 2022. *US Fed News*. Accessed via Nexis Uni.

5. *Methods and systems for automated software natural language documentation* Feb. 6, 2024. *Targeted News Service*. Accessed via Nexis Uni.

work on endangered languages and the automated creation of natural language materials in programming contexts. This highlights how the term has relevance across distinct professional domains — and how these may compete in discourse when it comes to the question which meaning takes the dominant role in media representation.

Topic 2. The top terms are *AI, bridge, data, inspection, agency, include, propose, policy, risk, system, information, government, report, FHWA* and *provide*. These keywords suggest that language documentation often occurs in a context dealing with technologies and governmental or industrial policies, including the corresponding authorities. The most frequent collocation from the topic's subcorpus is the bigram *plain-language documentation* ($\lambda = 9.35$, $z = 10.79$). To illustrate the topic's contexts, compare the following excerpts, with *language documentation* meaning something entirely different in each example. For instance, several reports on Australia's Housing Industry Association (HIA) describe the perks of plain language contracts for new housing in (6). Investigative stories pick up on non-English reports and refer to them as language documentation as exemplified in (7). And news coverage on guiding frameworks for AI transparency mention the blueprint for an AI Bill of Rights, shown in (8).

(6) HIA's plain language documentation [make] agreements between customers and builders easier to understand.[6]

(7) Drawing upon open-source Chinese language documentation, […] the report […] identifies some 83 well-known consumer brands linked to factories where Muslim minorities are thought to be working in forced labour conditions across China.[7]

(8) Designers of an AI system should provide generally accessible plain language documentation with clear descriptions of the overall system and the role that AI plays.[8]

In this topic, language documentation is not used in the sense of documentary linguistics, but instead refers to standardised, often policy-driven or technical writing designed to inform and regulate. The corresponding materials are set to ensure clarity, transparency, and compliance in governmental, legal, and technological domains, meaning this type of language documentation is entirely

6. *Package to provide service* Aug. 13, 1999. *Real Estate News*. Accessed via Nexis Uni.

7. Walden, Max. Mar. 2, 2020. "Apple, Nike and other major companies implicated in Muslim forced labour in China". *Asia Pacific News*. Accessed via Nexis Uni.

8. O'Connor, Cozen. May 15, 2023. "The Biden Administration's blueprint for an AI Bill of Rights". *Newstex Blogs*. Accessed via Nexis Uni.

detached from documentary linguistics. Hence, there is another competing type of language documentation in media representation.

Topic 3. The top terms are *fund, program, native, health, Indian, project, Alaska, service, process, include, provide, license, tribal, application* and *federal*. The most frequent collocation from the topic's subcorpus is the multi-word expression *in Alaska native language documentation* ($\lambda = 12.32$, $z = 2.03$). This is due to a large body of news coverage reporting on translations into Yup'ik and Inupiaq for certain official paperworks to apply for governmental funding, citing a linguistics professor specialising in these Alaska Native languages, as exemplified in (9). Other sources provide transcriptions of the Public Witness Hearing specifically for American Indian and Alaska Native Programs, mentioning the endangered Muscogean language that is spoken by the Seminole and Muscogee Creek Nations in Oklahoma, and the Seminole Tribe of Florida, compare (10).

(9) [He] has about three decades experience in Alaska Native language documentation and revitalization.[9]

(10) [The goal of the Spoken Creek Documentation Project] is to produce 24 hours of audiovisual recordings, transcriptions, and translations into English and train students in language documentation methods.[10]

In this topic, language documentation is tied to sociopolitical efforts aimed at the preservation and revitalisation of Indigenous languages, particularly within federally funded programmes. The term refers not just to academic activity, but to applied, community-driven projects that ensure linguistic and cultural accessibility in governmental and public services.

Topic 4. The top terms are *language, Indigenous, speak, community, people, speaker, native, English, learn, student, linguistic, school, culture, university* and *endanger*. These keywords suggest a strong focus on language documentation in the context of applied linguistics and the educational realm. This assumption is backed by the analyses of the topic's subcorpus: The most frequent collocation is the multi-word expression *language documentation and* ($\lambda = 3.33$, $z = 2.31$), which is complemented by the noun's *revitalisation efforts* and *revitalisation, conservation* as well as *description*. The second most frequent collocation is the multi-word expression *Indigenous language documentation education and* (though not significant, $\lambda = 9.07$, $z = 1.52$). Language documentation is presented in the context of conferences and festivals devoted to the preservation of Indigenous languages and

9. *FEMA fires group for nonsensical Alaska Native translations* Jan. 13, 2023. *Juneau Empire (Alaska)*. Accessed via Nexis Uni.

10. *H Approps hearing on American Indian and Alaska Native funding, morning session* May 16, 2017. *Financial Markets Regulatory Wire*. Accessed via Nexis Uni.

cultures, including the relevance it has to educate people about dying languages. For instance, a news release reports on a festival that features the Hawaiian language, in this context picking up on the relevance of language documentation (11). Another report is devoted to linguistics students and professors who learned about protests against the Dakota Access Pipeline and the attempt to revitalise the Lakota language, citing a language documentation expert (12). Meanwhile, another news release writes that the Language Conservancy unveiled the world's largest collection of Native American language learning materials, and that the nonprofit organisation will host a conference (13). Several outlets report on this conference, stating that it was set to educate stakeholders and urge immediate action to save Indigenous languages from extinction.

(11) [Case studies will highlight] the important role that language documentation and revitalization plays in sustaining cultural heritage and traditions.[11]

(12) Martin, who specializes in language documentation and revitalization and native languages of the American South, said that the pipeline controversy is connected to [...] the extinction of languages.[12]

(13) [The conference] brings together Indigenous First Language speakers, language experts, linguists, scholars, and practitioners from around the world to discuss the pressing issue of Indigenous language documentation and revitalization.[13]

Hence, the topic highlights language documentation as a central concern in academic and applied contexts, with a clear emphasis on its function in preserving endangered languages and the cultural identities they embody. News coverage underscores the role of universities in this process, portraying them as institutions where students are trained and resources are developed to support revitalisation efforts.

Topic 5. The top terms are *language, university, linguistic, research, endanger, speak, project, NSF, community, document, speaker, people, national, cultural* and *archive*. These keywords likely present language documentation in the context of certain research projects, funding agencies, principal investigators as well as the actual studied languages. The most frequent collocations in the topic's subcor-

11. *Hawaiian language to be featured at Smithsonian Folklife Festival* June 18, 2013. *US Fed News*. Accessed via Nexis Uni.

12. Nicholas, Antonella. Oct. 4, 2016. "Standing Rock protestors speak on panel, advocate against Dakota Access Pipeline". *Flat Hat: College of William and Mary*. Accessed via Nexis Uni.

13. *The Language Conservancy unveils largest collection of Native American language learning materials in the world at United Nations forum* Apr. 18, 2023. *MENAFN.com*. Accessed via Nexis Uni.

pus are the bigrams *and language documentation* ($\lambda = 1.87$, $z = 2.65$) and *on language documentation* ($\lambda = 2.01$, $z = 2.30$), backing this assumption. For instance, numerous news services report that the National Endowment for the Humanities (NEH) and the National Science Foundation (NSF) allocated a total of $ 4.5 million to document languages threatened with extinction, listing all researchers who received fellowships, institutional grants, and awards, including their projects as exemplified in (14). Similarly, there are further news reporting that a faculty member has earned funding to study and preserve Quechua, with the linguist explaining the relevance of their language documentation project in (15). Another outlet reports that a linguist has released a detailed resource on the Nama language of Southern New Guinea, providing a short biography as shown in (16).

(14) Jonathan Amith of Gettysburg College leads a Nahuatl language documentation project in Mexico, where the researchers are particularly interested in developing a grammar.[14]

(15) Additionally, this project will produce valuable resources for both academic research and language documentation and education, benefiting scholars and the communities involved.[15]

(16) His research has primarily focused on language contact, pidgins, creoles, and new dialects, and recently on language documentation.[16]

Hence, the topic clearly portrays language documentation as a structured and institutionally embedded research activity, describing all of its aspects: The articles encompass concrete projects, name principal investigators, and mention the endangered languages under investigation, while also highlighting the role of funding agencies and academic platforms in supporting and disseminating this work.

Topic 6. The top terms are *bank, offer, company, tender, service, market, financial, business, million, increase, share, billion, include, product* and *customer*. These keywords point to an economic context, suggesting that articles centre around economic affairs, particularly in the domains of financial services, business operations, and market activity. In the topic's subcorpus, typical bigrams are *English-language documentation* ($\lambda = 7.47$, $z = 23.87$), *Spanish-language documentation* ($\lambda = 5.60$, $z = 11.05$) and *Japanese-language documentation* ($\lambda = 11.11$, $z = 7.79$). The

14. *NEH and NSF Award $ 4.5 Million to Preserve Languages Threatened with Extinction* Aug. 9, 2012. *States News Service*. Accessed via Nexis Uni.

15. *FSU linguist awarded National Science Foundation grant to study and preserve Indigenous Quechua language* Oct. 15, 2024. *States News Service*. Accessed via Nexis Uni.

16. Berg, Tabitha. July 30, 2024. "Author Linguist Jeff Siegel Publishes Comprehensive 'Nama Dictionary' – A Detailed Resource on the Nama Language of Southern New Guinea". *eNewsChannels*. Accessed via Nexis Uni.

KWIC analysis shows that, although not significant, there are numerous additional modifiers of *language documentation* such as *Chinese, French, Hungarian* and *Ukrainian*. As regards examples, one news wire states that the Commercial Bank of Dubai has launched its TianLong comprehensive Chinese banking service, providing details as in (17). Another outlet reports on a tender, stating that the contractor will have to integrate a new future-proof language documentation system into the client's higher-level IT infrastructure and explaining further details as shown in (18). A tender is also discussed in another news wire's report, specifying that all procedures relating to the tender offer are to be conducted in Japanese, compare (19):

(17) The TianLong service includes Chinese language documentation, tailor-made products and services.[17]

(18) The language documentation system consists of hardware and standard software components.[18]

(19) If […] there is any inconsistency between the English language documentation and the Japanese language documentation, the Japanese language documentation will prevail.[19]

All told, within this topic, language documentation is not used in the sense of descriptive linguistics. Instead, it designates multilingual documentation used in banking, tenders, and customer-facing services and therefore textual materials such as contracts, product descriptions, or service manuals that are produced as part of institutional or commercial processes. Hence, the topic shows that there is yet another representation in public discourse competing for the dominant meaning of language documentation.

3.2 Topic trends over time

Having examined the organisational structure of the corpus in terms of topics and thus the content structure of the media discourse, we now move on to investigate diachronic trends in topic prevalence. To do so, we computed the average topic probabilities per publication date and visualise general trends using monthly aver-

17. *CBD launches comprehensive Chinese banking services* Dec. 19, 2012. *Islamic Finance News.* Accessed via Nexis Uni.

18. *Re-invest language documentation system* May 24, 2023. *Tenders Monitor Africa-Asia.* Accessed via Nexis Uni.

19. *MUFG Bank, Ltd. UK regulatory announcement: Tender offer 38 + 278* Dec. 27, 2024. *Market News Publishing.* Accessed via Nexis Uni.

ages. This allows for an interpretation of overall topical salience and long-term developments. Figure 2 illustrates the probability trends for each of the six topics during our investigated 25-year period, highlighting when certain themes gained or lost prominence. The x-axis represents the chronological progression of articles, while the y-axis displays the topic probability, capturing fluctuations in topic prevalence.

The figure is to be read as follows. A value of 1 shows that on average, this topic was very dominant in all articles published during a given point in time, meaning the topic was usually the main topic of the articles in that specific month. In turn, a value of 0 expresses that this topic is not attested in the period in question, meaning that none of the articles (or very few) were assigned to this topic in a month — at least not with any relevant probability. While we are less concerned with probabilities of single months, we are more interested in uncovering general trends and strengths for the six topics. A visual inspection reveals that topics 4 and 5 both show a continuous increase in average probability from 2000 onwards, remaining relatively stable as of 2010 and 2015, respectively. In addition, there are repeated spikes reaching values above 0.75. This pattern hints to a continuous presence and relevance of these thematic fields in the discourse over time. Similarly, although overall average probability is lower, the same trend is in principle attested for topic 1. Upticks are less pronounced, though. Moving on to topic 6, probabilities are relatively high around 2000, before the topic exhibits a marked decline, indicating a reduced salience in the discourse on average. However, high volatility as demonstrated by the repeated spikes can be observed as well, rising to values of 0.9 or above, suggesting that certain events were picked up on by news coverage, making it the dominant theme in discourse. Topic 2, meanwhile, exhibits lower average probabilities, although it also shows a few upticks when the topic was dominant. As regards topic 3, average probabilities are very low across the entire period, and only very few spikes are observed, meaning this topic is barely relevant for the media discourse.

To estimate how prominently each topic features in the overall dataset, independent of exclusive topic assignment, the aggregated document-topic probabilities (gamma values) can be used to reflect their relative prevalence. Drawing on this approach, the proportions show that across the corpus, topic 5 accounts for the largest share of content with 26.3%, followed by topic 4 (21.8%) and topic 1 (17.6%). Topic 6 (16.9%) and topic 2 (10.9%) follow with moderate representation, whereas topic 3 is the least dominant, covering only 6.3% of the overall topic distribution.

To recall, this distribution is a latent representation compiled by an algorithm and therefore to be considered a heuristic view to approach the types of representation of language documentation in the media. However, pairing this outcome

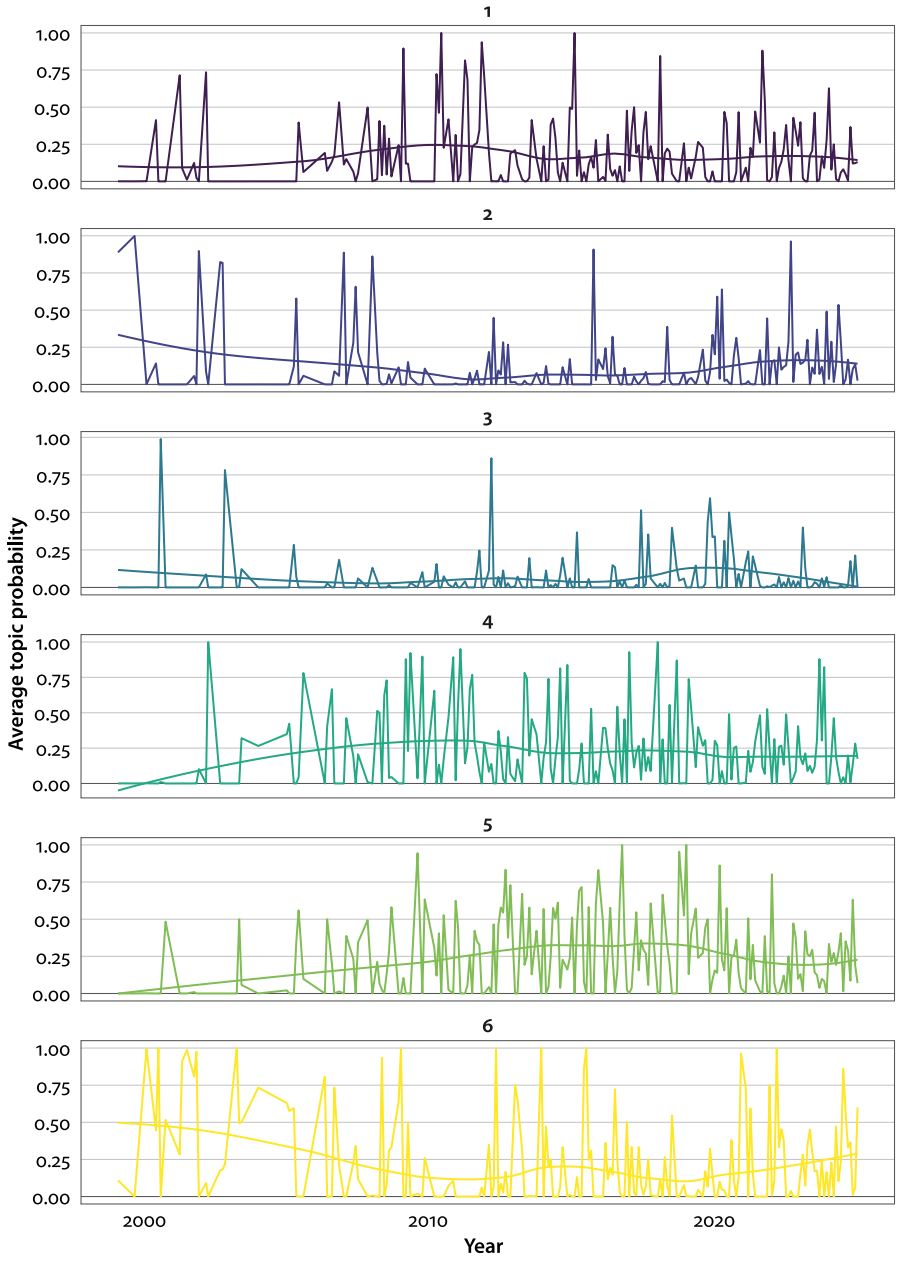


Figure 2. Evolution of different topics over time

with the results in the prior section, we are able to provide insights into the different ways of how language documentation is represented and framed in the media, quantifying the respective semantic profiles of the term. Compare the overview in Table 1.

Table 1. Media characterisations of language documentation (LD) by topic corresponding to documentary linguistics

Topic	Share	Match	Media characterisation
5	26.3%	✓	LD as research activity: Media mention concrete projects, PIs, languages under investigation, role of funding agencies and academic platforms
4	21.8%	✓	Function of LD: Preserving endangered languages and cultural identities; universities train students and develop resources to support revitalisation efforts
1	17.6%	✓/✗	LD as research field, presented in the context of portrayed experts; LD as technical writing in programming contexts
6	16.9%	✗	LD as textual materials such as contracts, product descriptions, or service manuals
2	10.9%	✗	LD as standardised, technical writing
3	6.3%	✓	LD as academic activity, tied to sociopolitical efforts aimed at the preservation and revitalisation of languages

The results come down to the finding that there are competing meanings of *language documentation* in the media. Although the dominant characterisation refers to the linguistic inquiry and preservation of linguistic primary data, including their analyses, almost 30% of mentions in public media discourse evoke a totally different connotation, referring to contracts, product descriptions, or service manuals instead. As topic 1 includes two types of meanings, which does not allow to disentangle how often the research field and technical writing is relevant, the number is likely to be higher, ranging up to 45% of mentions that do not refer to the academic discipline.

4. Discussion

The present pilot study posed the research question as to how language documentation is represented in the media. We now offer a discussion of our findings presented in the prior section.

4.1 Misrepresentation and gap between media and academia

As regards the Public Understanding of Science, the results show that language documentation is indeed not simply portrayed in the media in the same fashion as could be expected based on the academic perspective. Instead, language documentation appears in four topics, which are framed differently, respectively. Hence, the results support our initial assumption of the discursive construction, showing that the way language documentation is mentioned in the media and how it is embedded in news coverage affects its representation. The outcome therefore adds to prior findings on other academic disciplines, which demonstrated that media contribute to shaping the scholarly field. This is best evidenced if the media's discursive construction is compared with Himmelmann's (2006) definition of language documentation. The diverging perspectives concern several aspects, with media (i) highlighting certain characteristics, (ii) overemphasising certain facets, and (iii) ignoring, or simply not mentioning, relevant strands. First, the preservation of languages is obviously something that is a dominant theme in the media. Second, language documentation is fairly often framed in terms of revitalisation efforts and the preservation of cultural identities, meaning the mere results of language documentation are exaggerated or misreported. Third, media actually never mention *how* language documentation or the preservation of languages is carried out or how the goal of revitalising a language is achieved. The compilation of corpora, the need to annotate these primary data and add metadata is not mentioned. And the fact that language corpora and repositories aid in language maintenance do not appear in news coverage. There is also no distinction provided between language documentation, i.e. collecting data, and providing a descriptive analysis of these data, i.e. language description. Hence, there is a strong bias in media representation concerning language documentation, promoting some, while obscuring other aspects of the discipline. This finding points to a clear gap and misrepresentation of language documentation between media and academia.

4.2 Psychological implications

The TM approach also has some psychological implications. Firstly, the analysis shows that the relative importance of certain topics in news coverage varies over time, suggesting that topic distribution is not fixed. This observation may be interpreted in light of agenda-setting processes by media professionals, although such an interpretation goes beyond the evidence provided by the corpus itself. Mass media can assume the function of attention allocation by enabling their consumers to deduce the societal and personal significance of a topic based on the

amount and prominence of information presented in mass media (McCombs & Shaw 1972). Conversely, this means that a (temporary) lack of visible reporting in the mass media on topics such as language documentation cannot generate corresponding attention from the public. In this respect, it is up to the people involved in the research to place their research work in the mass media.

Secondly, this also directly results in the necessity of not understanding mass media as a homogeneous mass of publication organs. In particular, the increasing importance of social media for science communication and the significantly greater direct participation of the public in this context means that scientists are faced with the challenge of presenting their topics on suitable platforms in an appropriate form (Foell 2021). Also, the units of analysis vary depending on the social media platform and sometimes differ significantly from the newspaper articles analysed here as examples (e.g. videos on YouTube, texts in the form of comments, audio tracks in the context of podcasts). However, attention must be paid to the characteristics of the specific media source. For example, visible competence of the source is important for the consumers' perception of source credibility and message credibility as well as consumers' news engagement, such as giving the news a "like" on social media platforms, leaving a comment, or talking to someone about the articles (e.g. Zimmermann et al. 2023). News engagement is therefore particularly important with regard to the dissemination of topics, such as language documentation, within the public and thus their visibility and potential impact.

Thirdly, it must be taken into account that not all sections of the public can be reached by all mass media. In addition to consumers' individual interests (e.g. Kaspar & Fuchs 2021), the general availability of certain media and the individual's ability to consume these media also play a role. Inter-individual differences in access to media are well documented (e.g. Wang et al. 2025). In this respect, when the question arises as to whether a topic such as language documentation can cause intended effects in the public through media coverage, it is not enough to analyse only the side of the coverage itself, but a target group analysis and their media usage behaviour seem obligatory.

4.3 Future avenues

As follows from the discussion, there is a range of potential directions for future research that could be pursued.

First, future research could aim to replicate the present study using a different dataset, focusing, e.g. on select countries or regions. A balanced range of news coverage could include outlets with varying degrees of relevance and influence for experts and lay people, respectively, ranging from major national newspapers

to regional publications and online platforms. Such an approach would allow for a deeper understanding of public discourse. In addition, it would be valuable to rerun the analysis, zooming in on shorter time windows. Drawing on a qualitative approach, such an analysis should pay attention to the observed peaks and shifts in topic prevalence, and how these relate to external developments such as policy reforms, technological innovations, or key sociopolitical events. Establishing such connections would strengthen the interpretation of trends and patterns within the data.

Second, following up on the latter point, continued research could be complemented by a differentiated breakdown of different suggested time periods. In addition, two further dimensions for in-depth qualitative research present themselves: Deep qualitative multimodal discourse analyses of selected public representations of language documentation could help us better understand their function in a context of open and mass media science communication. Beyond these existing sources, audience reactions could be studied both through existing accounts given by audience members, and through directed audience interviews. In each of these ways, our current assumptions about recent media history may either be tested for sustainability or eventually corrected as media conditions impact the representation and, eventually, the public understanding of language documentation.

Third, future work could explore the role of social media in shaping or amplifying narratives around linguistic diversity. Comparing traditional media coverage with user-generated content may reveal differing framings or emergent grassroots discourses.

Finally, it must also be noted that the approach pursued here, despite its strength in aggregating many sources and being able to categorise and chronologically classify certain thematic highlights, remains interpretative and speculative to some extent. The keywords identified, which define a thematic area, cannot be used to fully reconstruct the original arguments and the connections between the key constructs presented in the original sources. In addition, different media sources report on the same events in very different ways and this different framing (Cacciatore et al. 2016; Vliegenthart 2012) can lead to different effects among news consumers (e.g. Lecheler & Vreese 2012). Storytelling has a significant function for psychological processing — including interest in the topic and one's own behavioural intention — both in classic text-based (e.g. Kaspar et al. 2016) and more immersive digital media formats (e.g. Zimmermann et al. 2025). In this respect, it would be highly desirable to pursue the analytical approach of media reporting presented here not only in the sense of classic agenda setting strategies, but also to supplement it with aggregating analyses of media framing. This step would mean to not only investigate *that* different topics are reported on and

certain types of representation of language documentation are found, but also explain *how* the topics are reported on and why various narratives and framings gain traction over time.

4.4 A call for engagement

In Section 1.1, we have picked up on science communication, summarised the relevance of public trust in science, and explained why research outcomes have to be meaningful to a wider audience. In light of the present study's results, which show to some extent a misrepresentation of the discipline's activities and goals in public media, the necessity for researchers in the field of language documentation to engage in broader outreach efforts becomes ever more pressing. Communicating research results and explaining what language documentation does, which languages are investigated, which methods are used and how this plays out on the level of language preservation and revitalisation as well as how it affects cultural identities is not trivial. Instead, seeing that language documentation is only in slightly more than half of all mentions represented in the media as it is understood in academia, such outreach efforts actively contribute to steering and reifying the dominant discursive understanding of the academic discipline, which in turn affects how language documentation is constructed in public belief and laypeople's attitudes.

5. Conclusion

Driven by the research questions of how documentary linguistics is represented in the media, this pilot study examined a corpus of English-language newspaper articles to identify dominant topics in public discourse in which language documentation is embedded.

Rooting our study theoretically in the Public Understanding of Science framework and a rough overview of recent shifts in public science communication, we used Topic Modelling and a corpus-assisted discourse approach, specifically keywords in context and collocation analysis, to investigate in which contexts the term *language documentation* is presented and framed in the media. We identified six topics, at least three of which portray the academic discipline, presenting language documentation mainly as a research and academic activity. Overall, a good one in two mentions in the media belong to these three topics. Nevertheless, media tend to highlight certain characteristics of the discipline, overemphasise certain facets, and sometimes ignore relevant strands of language documentation. Two topics show that a competing meaning of language documentation circulates

in the media discourse, referring to technical writing as well as textual materials such as contracts, product descriptions, and service manuals.

The findings are relevant for researchers working in the field of language documentation as they highlight the necessity to actively communicate and report on the discipline and its outcome in an effort to ensure it takes the dominant meaning in media discourse. This, in turn, is motivated by the relevance to steer public engagement and awareness, which are directly linked to fostering public trust, securing funding, and ensuring broad-based support for linguistic research.

Data availability

Since the corpus analysed in this study consists of newspaper articles obtained via Nexis Uni, the full texts cannot be shared publicly due to copyright restrictions. However, metadata, the articles' gamma values and the output as well as processing and analysis scripts are available at the Open Science Framework [https://osf.io/k4hdj/].

References

Abdelrazek, Aly, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. "Topic Modeling Algorithms and Applications: A Survey." *Information Systems* 112: 102131.

Ampollini, Ilaria, and M. Bucchi. 2019. "When Public Discourse Mirrors Academic Debate: Research Integrity in the Media." *Science and Engineering Ethics* 26: 451–474.

Aranda, Ana M., Kathrin Sele, Helen Etchanchu, Jonne Y. Guyt, and Eero Vaara. 2021. "From Big Data to Rich Theory: Integrating Critical Discourse Analysis with Structural Topic Modeling." *European Management Review* 18 (3): 197–214.

Attenborough, Frederick Thomas. 2013. "Discourse Analysis and Sexualisation: A Study of Scientists in the Media." *Critical Discourse Studies* 10 (2): 223–236.

Bednarek, Monika. 2024. "Topic Modelling in Corpus-based Discourse Analysis: Uses and Critiques." *Discourse Studies* 27, online first.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. . https://quanteda.io

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84.

Bonfadelli, Heinz. 2017. "Handlungstheoretische Perspektiven auf die Wissenschaftskommunikation." In *Forschungsfeld Wissenschaftskommunikation*, ed. by Heinz Bonfadelli, Birte Fähnrich, Corinna Lüthje, Jutta Milde, Markus Rhomberg, and Mike S. Schäfer, 83–105. Wiesbaden: Springer Fachmedien Wiesbaden.

Bourdieu, Pierre. 1996. *Sur la télévision*. Paris: Liber-Raison d'agir.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication and Society* 15 (5): 662–679.

Brookes, Gavin, and Tony McEnery. 2019. "The Utility of Topic Modelling for Discourse Studies: A Critical Evaluation." *Discourse Studies* 21 (1): 3–21.

Brossard, Dominique, and Bruce V. Lewenstein. 2010. "A Critical Appraisal of Models of Public Understanding of Science: Using Practice to Inform Theory." In *Communicating Science: New Agendas in Communication*, ed. by LeeAnn Kahlor, and Patricia Stout, 11–39. New York: Routledge.

Buzova, Daniela, Silvia Sanz-Blas, and Cristina Santos-Rojo. 2024. "Disruptive Innovation Discourse: Are Academic Research and News Media on the Same Page?" *Technology Analysis and Strategic Management* 36 (7): 1353–1368.

Caballe, Anna, and Martino Bardelli. 2021. "Building Blocks of Virtuous Science Communication: Grant Funding, Policy Making, and Public Engagement." *DNA and Cell Biology* 41: 6–10.

Cacciatore, Michael A., Dietram A. Scheufele, and Shanto Iyengar. 2016. "The End of Framing as we Know it… and the Future of Media Effects." *Mass Communication and Society* 19 (1): 7–23.

Cascone, Kim. 2000. "The Aesthetics of Failure: 'Post-Digital' Tendencies in Contemporary Computer Music." *Computer Music Journal* 24 (4): 12–18.

Cho, Hae-Wol. 2019. "Topic Modeling." *Osong Public Health and Research Perspectives* 10: 115–116.

Churchill, Rob, and Lisa Singh. 2022. "The Evolution of Topic Modeling." *ACM Computing Surveys* 54 (10s): 215.

Czaykowska-Higgins, Ewa. 2018. "Reflections on Ethics: Re-humanizing Linguistics, Building Relationships across Difference". In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 110–121. University of Hawai'i Press. http://hdl.handle.net/10125/24813

Dijck, José van. 2003. "After the 'Two Cultures': Toward a '(Multi)cultural' Practice of Science Communication." *Science Communication* 25 (2): 177–190.

Durant, John R., Geoffrey A. Evans, and Geoffrey P. Thomas. 1989. "The Public Understanding of Science". *Nature* 340: 11–14.

Eagleman, David M. 2013. "Why Public Dissemination of Science Matters: A Manifesto." *Journal of Neuroscience* 33 (30): 12147–12149.

Farr, Robert M. 1993. "Common Sense, Science and Social Representations." *Public Understanding of Science* 2 (3): 189–204.

Foell, Jens. 2021. "Social Media Science Communication is a Nonstop Academic Conference for all." *Nature Human Behaviour* 5 (7): 812.

Fraser, Nancy. 2007. "Special Section: Transnational Public Sphere: Transnationalizing the Public Sphere: On the Legitimacy and Efficacy of Public Opinion in a Post-Westphalian World." *Theory, Culture and Society* 24 (4): 7–30.

Gilboa, Eytan. 2005. "The CNN Effect: The Search for a Communication Theory of International Relations." *Political Communication* 22 (1): 27–44.

Gregory, Jane, and Stephen W. Miller. 1998. *Science in Public: Communication, Culture and Credibility*. New York: Plenum Trade.

Grün, Bettina, and Kurt Hornik. 2011. "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40 (13): 1–30.

Grün, Bettina, and Kurt Hornik. 2024. *topicmodels: Topic Models. R Package* version 0.2–17. https://CRAN.R-project.org/package=topicmodels

Habermas, Jürgen. 2022. *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*. Frankfurt am Main: Suhrkamp.

doi Happer, Catherine, and Greg Philo. 2013. "The Role of the Media in the Construction of Public Belief and Social Change." *Journal of Social and Political Psychology* 1 (1): 321–336.

doi Hilgartner, Stephen. 1990. "The Dominant View of Popularization: Conceptual Problems, Political Uses." *Social Studies of Science* 20 (3): 519–539.

doi Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–196.

doi Himmelmann, Nikolaus P. 2006. "Language Documentation: What is it and What is it Good for?" In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel 1–30. Berlin and New York: De Gruyter Mouton.

Holton, Gary, and Mandana Seyfeddinipur. 2018. "Reflections on Funding to Support Documentary Linguistics." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 100–109. University of Hawai'i Press. http://hdl.handle.net/10125/24812

doi Horst, Maja. 2013. "A Field of Expertise, the Organization, or Science Itself? Scientists' Perception of Representing Research in Public Communication." *Science Communication* 35 (6): 758–779.

Imhof, Kurt. 2011. *Die Krise der Öffentlichkeit. Kommunikation und Medien als Faktoren des sozialen Wandels*. Frankfurt am Main and New York: Campus.

doi Jacobs, Thomas, and Robin Tschötschel. 2019. "Topic Models Meet Discourse Analysis: A Quantitative Tool for a Qualitative Approach." *International Journal of Social Research Methodology* 22 (5): 469–485.

Jäger, Siegfried. 2012. *Kritische Diskursanalyse. Eine Einführung*. 6th ed. Münster: Unrast.

doi Jarren, Otfried, Kurt Imhof, and Roger Blum (eds). 2000. *Zerfall der Öffentlichkeit?* Wiesbaden: Springer.

doi Jonker, Hans, and Florian Vanlee. 2024. "Linking Science with Media and Policy: The Case of Academics in Flanders, Belgium." *Quantitative Science Studies* 5 (3): 556–572.

doi Kaspar, Kai, and Lisa Anna Marie Fuchs. 2021. "Who Likes What Kind of News? The Relationship Between Characteristics of Media Consumers and News Interest." *SAGE Open* 11 (1): 21582440211003089.

doi Kaspar, Kai, Daniel Zimmermann, and Anne-Kathrin Wilbers. 2016. "Thrilling News Revisited: The Role of Suspense for the Enjoyment of News Stories." *Frontiers in Psychology* 7: 1913.

doi Lecheler, Sophie, and Claes H. de Vreese. 2012. "News Framing and Public Opinion: A Mediation Analysis of Framing Effects on Political Attitudes." *Journalism and Mass Communication Quarterly* 89 (2): 185–204.

Linn, Mary S. 2018. "Reflections on Public Awareness." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 66–74. University of Hawai'i Press. http://hdl.handle.net/10125/24809

doi Luzón, María José, and Carmen Pérez-Llantada (eds). 2019. *Science Communication on the Internet*. Amsterdam and Philadelphia: John Benjamins.

Mach, Katharine J., Raúl Salas Reyes, Brian Pentz, Jennifer Taylor, Clarissa A. Costa, Sandip G. Cruz, Kerronia E. Thomas, James C. Arnott, Rosalind Donald, Kripa Jagannathan, Christine J. Kirchhoff, Laura C. Rosella, and Nicole Klenk. 2021. "News Media Coverage of COVID-19 Public Health and Policy Information." *Humanities and Social Sciences Communications* 8: 220.

McCombs, Maxwell E., and Donald L. Shaw. 1972. "The Agenda-Setting Function of Mass Media." *The Public Opinion Quarterly* 36 (2): 176–187. https://www.jstor.org/stable/2747787.

Millar, Robin, and Brian Wynne. 1988. "Public Understanding of Science: From Contents to Processes." *International Journal of Science Education* 10 (4): 388–398.

Mitchell, Madeline, and Merryn McKinnon. 2019. "'Human' or 'Objective' Faces of Science? Gender Stereotypes and the Representation of Scientists in the Media." *Public Understanding of Science* 28 (2): 177–190.

Pepperell, Robert, and Michael Punt. 2000. *The Postdigital Membrane. Imagination, Technology and Desire*. Bristol: Intellect.

Peters, Hans Peter. 2013. "Gap Between Science and Media Revisited: Scientists as Public Communicators." *Proceedings of the National Academy of Sciences* 110 (supplement_3): 14102–14109.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Rinker, Tyler W. 2018. *textstem: Tools for Stemming and Lemmatizing Text*. Version 0.1.4. Buffalo, New York. http://github.com/trinker/textstem

Schäfer, Mike S. 2012. "Taking Stock: A Meta-analysis of Studies on the Media's Coverage of Science." *Public Understanding of Science* 21 (6): 650–663.

Silge, Julia, and David Robinson. 2016. "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *Journal of Open Source Software* 1 (3): 37.

Snow, Charles Percy. 1993. *The Two Cultures*. Canto edition. With an introduction by Stefan Collini. Originally published 1959. Cambridge: Cambridge University Press.

Sumner, Petroc, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A. Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D. Chambers. 2014. "The Association between Exaggeration in Health Related Science News and Academic Press Releases: Retrospective Observational Study." *BMJ* 349.

Vayansky, Ike, and Sathish A. P. Kumar. 2020. "A Review of Topic Modeling Methods." *Information Systems* 94: 101582.

Vliegenthart, Rens. 2012. "Framing in Mass Communication Research: An Overview and Assessment." *Sociology Compass* 6 (12): 937–948.

Wang, Kun, Xiayu Summer Chen, Yanjun Dong, Karla G Sanabria Véaz, and Danan Gu. 2025. "Bridging the Access Gap: A Decade of Narrowing the Digital Divide for Hispanic Older Adults in the United States." *Journal of Aging and Health* 37 (3–4): 182–191.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. . https://ggplot2.tidyverse.org

Wickham, Hadley. 2024. *rvest: Easily Harvest (Scrape) Web Pages*. R package version 1.0.4. https://CRAN.R-project.org/package=rvest

Wynne, Brain. 1992. "Public Understanding of Science Research: New Horizons or Hall of Mirrors?" *Public Understanding of Science* 1 (1): 37–43.

Yearley, Steven. 1994. "Understanding Science from the Perspective of the Sociology of Scientific Knowledge: An Overview." *Public Understanding of Science* 3 (3): 245–258.

Yin, Yian, Yuxiao Dong, Kuansan Wang, Dashun Wang, and Benjamin F. Jones. 2021. "Science as a Public Good: Public Use and Funding of Science." *IO: Productivity*.

Yin, Yian, Yuxiao Dong, Kuansan Wang, Dashun Wang, and Benjamin F. Jones. 2022. "Public Use and Public Funding of Science." *Nature Human Behaviour* 6: 1344–1350.

Yu, Jingqi, Catherine Yeung, and Dilip Soman. 2025. "Are Media Reports of Published Research an Accurate Representation of the Research?" *Behavioural Public Policy* 2025: 1–21.

Zimmermann, Daniel, Asina Klee, and Kai Kaspar. 2023. "Political News on Instagram: Influencer versus Traditional Magazine and the Role of their Expertise in Consumers' Credibility Perceptions and News Engagement." *Frontiers in Psychology* 14: 1257994.

Zimmermann, Daniel, Paulina Wolf, and Kai Kaspar. 2025. "Virtual Reality versus Classic Presentations of Mass Media Campaigns: Effectiveness and Psychological Mechanisms Using the Example of Environmental Protection." *Computers in Human Behavior* 168: 108643.

PART II

Documentary corpora

CHAPTER 8

Taking language use seriously

On the importance of multimodality and interaction in documentary practices

Mandana Seyfeddinipur[1], Pamela Perniss[2]
& Claudia Wegener[2]
[1] Berlin-Brandenburg Academy of Sciences and Humanities
[2] University of Cologne

The aim of language documentation is to provide a comprehensive record of the observable linguistic behaviour and practices of a language community to be used now and in the future. The endeavour goes beyond the mere description of the linguistic system. To serve as a multipurpose record, the documentary record must capture language use in its fullest form — as a fundamentally multimodal and interactional phenomenon. This means that video recording is essential to the collection of data in language documentation efforts, and that the description and analysis of data must ultimately extend to and integrate multimodal and interactive components of language use. In this chapter, we outline the implications for language documentation from such a conception of language.

Keywords: language use, interaction, multimodality, gesture, sign language, language documentation

1. Introduction

As a subdiscipline of linguistics, the field of language documentation aims to compile a comprehensive and multipurpose record of natural languages (Himmelmann 2006). What exactly that is, however, is not straightforward, and in itself raises questions about what constitutes such a record of language. As Himmelmann (2006) notes, the definition of the word *language* creates the first potential problem in deciding what data to collect when documenting a language. The theoretical conception has critical practical implications in terms of methodologies and how we conceive of the criteria for what constitutes a comprehensive and multipurpose record of a language. Himmelmann advises to cast the net as widely as possible,

https://doi.org/10.1075/slcs.240.08sey
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

including "all aspects of the set of interrelated phenomena commonly called *a language*" (2006: 2), and referring mainly to the inclusion of different varieties, registers, as well as examples of spoken and written language. In this paper, we describe the theoretically grounded need for documenting language proceeding from a conception of language as a multimodal and interactional phenomenon. Crucially, this includes features of language use — notably, gestures — that have not been considered to be part of the formalizable system of language. After a short section outlining the aims of language documentation, we recapitulate some aspects of the historic development of the field of language documentation, showing how the conception of language and the technological possibilities of recording language have played a role in guiding ideas of what should constitute language documentation. We then argue for a fundamental theoretical conception of language as a multimodal and interactional phenomenon and provide illustrative examples of studies that show how important the consideration of multimodality is for a deeper understanding of language. We will then turn to what we see as implications for the field of language documentation following from the view of language as fundamentally multimodal and interactional.

2. Aim of language documentation

The basic idea in compiling a comprehensive and multipurpose record of a language (Himmelmann 1998, 2006, 2012; Woodbury 2003, 2011) is to record the way language is used in different contexts relevant to a community's language use ecology, such that the created record can be used for many different purposes, from language maintenance to revitalization, from linguistic analysis to the study of botany, religion, or history. Squaring the now available technological means of capturing communicative events with the prerequisites for scientific work and the potential interests and requirements of a language community, the ideal documentation, following Himmelmann (2012), would consist of "raw data" in the form of video recordings, including good audio quality, of all possible communicative events, and across community members and ages, in a way that shows the whole communicative situation and all participants. To allow for further research, as well as for maintenance and revitalization efforts (Hinton et al. 2018), "primary data" must be derived from the raw data in the form of transcripts (with segmentation on various levels: sound segments, words, intonation units) and translations. From here, further analysis in the form of "structural data" is possible, including glossing with an analysis of grammatical and lexical structure as well as linguistic and typological descriptions. This layer is not just important to researchers previously unfamiliar with the language, but also, for example,

for descendants of a language community trying to revitalize a language that no longer has native users.

The balance between the amounts of raw data primary data, and structural data that can be provided in practice is greatly dependent on the nature of individual documentation projects. Ideally, a comprehensive documentation project would comprise raw data of observed linguistic behaviour of different kinds, in all possible contexts and on all topics, as well as elicited and experimental data (Himmelmann 2012). Attaining this ideal is obviously offset by the realities on the ground, which must be practically and pragmatically resolved. In addition, to the extent that language documentation is often associated with situations of language endangerment, a fundamental premise of language documentation may be to record as much as possible, and even to delay analysis, before it is too late (Himmelmann 1998). The amount of raw and primary data necessary to allow recordings to be usable at a later stage, possibly when the language is no longer spoken, is a matter of debate, with some arguing that as much grammatical description as possible is necessary for usability (see Gippert et al. 2006; Evans 2008: 346–348; and Himmelmann 2012 in response to Evans 2008; Bowern 2015; Meakins et al. 2018). As such, the difference between documentary linguistics (centred on the relationships between raw data and primary data) and descriptive linguistics (centred on the relationships between primary data and structural data) is somewhat fluid. In short, a fully comprehensive and systematic documentary approach is concerned with the creation of a digital record of a language through recording, transcribing and annotating language use, as well as archiving of the data in a trusted repository and searchable on the basis of comprehensive metadata, and including detailed information on the setting, participants and the nature of their relationships (e.g. Gippert et al. 2006).

The most realistic and pragmatically feasible approach may be to think of documentary practice in terms of concentric circles: a wealth of raw data is recorded; some of it is transcribed and translated (primary data); some of it is glossed (structural data); and a little of it is annotated in more detail for specific research questions (structural data). The main take-away point is to find the right balance between what is needed for a specific research question at the time of recording, within the constraints of the specific research project and according to the wishes and needs of the specific language community, and to maximize potential based on what might be useful at some later stage, potentially by the same researcher or community, a member of a later generation of researchers, or descendants of the community interested in revitalization.

3. Some observations on the history of documentation

As a practice, language documentation was proposed, in large part, in response to the outcry among linguists about the dramatic increase of language endangerment and loss (Krauss 1992; Hale et al. 1992). The realization that without extensive documentation, a large number of languages will disappear with very little or no trace, and with them a huge part of data needed for language maintenance and revitalization as well as for a better understanding of human language and cognition, shifted the focus to the increased collection of raw data (i.e. recordings of observed linguistic behaviour). Building on a long history of the collection of word lists, grammatical sketches and texts — going back to Gottfried Leibniz, Wilhelm von Humboldt as well as Franz Boas (e.g. Duerr 2013; La Polla 2020) — as a means necessary to conduct linguistic research, Himmelmann's seminal paper in 1998 proposed a program for documenting languages, taking advantage of digital technologies allowing large amounts of data to be recorded, with the aim of compiling a comprehensive record of a language that allows as wide a range of uses as possible.

The development of and easy access to increasingly sophisticated recording technology, and crucially also the potential to store data in accessible repositories, enabled more and more in-depth and, importantly, reproducible and replicable analyses in linguistics. First, the invention of magnetic tape allowed recordings to be replayed over and over again, enabling researchers to analyse vocal features, intonation and prosody with more certainty and a more rigorous scientific approach. The advent and increasing accessibility of video-recording technology marked another transformative moment in the empirical study of linguistic behaviour. The video camera introduced the possibility of capturing the full multimodal ecology of human interaction. While ethnographic filmmaking and visual anthropology have used film as a research tool and as a tool for storytelling, with cinematic techniques to convey cultural narratives, since the beginning of the 20th century (MacDougall 1998; Henley 2020), linguistic anthropology (see Duranti 1997, 2005) as well as conversation analysis centred the recording on the creation of a record used for analysis (e.g. Mondada 2008; Parry 2010), corresponding to what is called raw data in language documentation.

The advent of video-recording technology allowed researchers to systematically observe and analyse embodied practices (e.g. gesture, gaze, facial expression, posture) that co-occur with speech and contribute fundamentally to meaning making. Importantly, video recording also made possible the systematic investigation and analysis of sign languages. It enabled a more holistic approach to discourse, one that recognizes that communication is distributed across multiple semiotic channels in both spoken and signed language use. Phenomena such as

the coordination of talk and bodily conduct in the coordination of turn taking and repairs, and the role of material environments in structuring interaction also became possible to record (see e.g. Kendon 1967, 1979; Goodwin & Goodwin 1986; Goodwin 1986, 2014; Streeck 2009). It is important to note, however, that the study of gesture and sign has a long history, going back long before video technology became available. Kendon (1983), for example, gives an overview of the history of observation-based descriptions of gesture dating back to antiquity. With the advent of digital technology, field linguists were able to collect large amounts of high-fidelity audio and video recordings of linguistic data in a compact, durable and easily reproducible format. For documentation, this means that substantial amounts of data can be recorded, language use can be documented, and the materials can be archived and preserved. The nature of archived recordings differs widely, however. In part, this is due to the obvious desideratum of recording as many different types of data as possible, including narratives, procedural texts, natural conversations, as well as elicited data. The nature of recordings is also, however, determined in part by the theoretical conception of what constitutes language, i.e. what is being documented. Much of language research has been biased by a view of language as a unimodal, monologic phenomenon, as just speech or just written text (Linell 2004; Seyfeddinipur 2012; Vigliocco et al. 2014; Perniss 2018), and this can sometimes still be observed in language documentation practice. What can effectively be called a written language bias (Linell 2004; Ong 1982) — encouraged as well by the focus on "langue" (Saussure 1916) or "competence" (Chomsky 1965, 1986) to the exclusion of "parole" or "performance" as the observable and actual use of language — has influenced descriptive and theoretical linguistics as well as language documentation.

An area where the written language bias and its effects in language documentation becomes apparent is that audio-only recordings of speech continue to constitute large parts of archived documentation material (Seyfeddinipur et al. 2019). While there are certainly situation-specific important reasons not to record video (e.g. taboos on capturing people images, community or speaker wishes), recording only audio when video is possible, can be detrimental. This can be the case for what is commonly referred to as procedural texts, narratives in which speakers describe to a novice, often the documenter, how they produce elements of material culture, e.g. how to make an arrow, prepare a traditional dish, build a boat, or weave a basket. These take place in co-presence, face-to-face, with speakers demonstrating how objects are manipulated and created. In the following example, an artist is explaining to a documenter how to weave a basket with a unique pattern used in a ritual through which a shaman receives her power. Translated the speech reads like this (excerpt from anonymized data): "You take this and put it here and then you go under and up again and hold that up so it can go

here. Then you put it here and put this there but hold, you wait here, hold and then push up." What can be observed is a demonstration (Clark & Gerrig 1990), which is characterized by the proliferation of deictic elements and reduced lexical specificity (Hanks 2005). The information distribution relies heavily on the visual channel, showing the steps of weaving, with the verbal production indexing the actions performed. The way the artist describes how to weave a basket is dependent on the face-to-face situation, where interlocutors can see each other, as well as the common ground between interlocutors. Hence, the artist is able to simply point to things and the interlocutor may simply nod, for example, to indicate understanding. With a recording limited to audio, this record of language use cannot even fulfil the purpose of reconstructing what was created and how it was created. Video provides the ideal tool to capture this genre, which is fundamental in language documentation. Even where the call for video recording is heeded (Himmelmann 1998, 2006; Schultze-Berndt 2015), a biased conceptualization of language in the methods used to record linguistic data in the field may be implicit. For example, when individual speakers are positioned alone in front of a camera (e.g. to relate a personal narrative) the dimensions of contingency, negotiation and multimodality of language are lost or retreat into the background, as the speaker is isolated from the environment where such narratives usually take place.

The production of transcripts from the raw data recordings is also influenced by a written language bias. For example, when orthographic transcriptions are produced, conventions from the written modality of the major contact language (e.g. commas, full stops, exclamation marks and question marks) are often used. While segmentation must necessarily occur, the imposition of these language-specific conventions onto transcripts of spoken discourse should not be uncritically imposed. It is important here to clearly define the role of each element in the given transcript and to document these (see Jefferson 2004 for such a description for conversation analytic transcripts). At the same time, features characteristic of spontaneous spoken language — such as pauses, fillers like *uhm* and *uh*, or false starts — may be omitted from transcripts (Linnell 2005:205). "Editing out" these elements gives the impression that natural speech does not include such phenomena, while they are actually highly informative indicators of processing difficulties or discourse structuring (Clark & Fox Tree 2002; Swerts 1998) or constitute conventional elements in and across languages (Pakendorf & Rose 2025). This is not to say that a conversation analytic transcription is required, but that the transcription conventions and decisions should be carefully reflected and documented in the documentary record (see Section 6.2).

It is important to be aware of and reflect on what is potentially missing or lost, and what this means for our understanding of language and the creation of a multipurpose record which can be used above and beyond our immediate interest. With

a bias toward a unimodal, monologic conception of language, the danger of ignoring these considerations looms larger. In order to be able to notice, analyse and describe multimodal negotiation and communicative practices between speakers, we need video recordings of interactions that show not only the speaker but also the interlocutor(s). We still know too little about how conversation is structured, and how recipient design influences the collaborative nature of conversation (Clark 1996), in different languages to fully understand what cross-linguistic patterns there are and what the range of variation is.

Finally, the recognition of sign languages as worthy of study and documentation was hampered by a unimodal conception of language as speech or written text (Vigliocco et al. 2014). For centuries, the primacy of speech as constitutive of language and the equation of speech with language and intellectual ability meant that sign languages were not recognized as languages and deaf people were considered incapable of fathoming and expressing complex and abstract content. This is reflected also in audism, a belief in the superiority of the ability to hear and use spoken language, which permeates discriminatory attitudes against deaf people and their cultures and languages (Humphries 1977).

4. Language use, multimodality and joint action

We advocate for a fundamental conception of language as a multimodal, interactional and negotiated phenomenon, and for the importance of incorporating this conception into documentary practice. Scholars from fields like sociology (e.g. Goffmann 1959), ethnomethodology (e.g. Garfinkel 1967), ethnography of speaking (e.g. Hymes 1962, 1964; Sherzer 1983), conversation analysis (e.g. Sacks et al. 1974; Jefferson 2004; Sacks 1992), sociolinguistics (Labov 1972; Gumperz & Hymes 1964), linguistic anthropology (e.g. Ochs 1988; Duranti 1997), and psychology (e.g. Clark 1996; Bavelas 2022), to name a few and some of their seminal publications, have conceptualized language as part and parcel of sociocultural structures and processes. Communication is viewed as joint construction by interactants, shaped by social and cultural contexts, and involving dynamic processes of meaning making. To this end, language is analysed as a multimodal system that integrates multiple modes of communication — such as speech, gestures, facial expressions, gaze, posture and spatial arrangements — to create, negotiate and convey meaning dynamically in social interaction. This reflects a shift from viewing language purely as an abstract code of arbitrary symbols to understanding language in its social, embodied, contextual and interactive dimensions, where various semiotic resources cooperate in the meaning-making process across linguistic, visual, gestural, auditory and spatial channels.

With a focus on language use as negotiated and multimodal, and as a form of joint action, we will be able to understand a wide range of dimensions of language as following rules and conventions above and beyond (our standard conception of) grammar. For example, rules that regulate conversation include the management of turn-taking (Sacks et al. 1974; Schegloff 1998) and repair (Schegloff et al. 1977; Schegloff 2009; Bolden 2011; Dingemanse & Enfield 2015; Dingemanse et al. 2015; Albert & de Ruiter 2018) and the use of backchannels like *mhmh*, head nods, or head shakes (Bangerter & Clark 2003; Schegloff 1982) as well as everyday social activities, for example inviting a friend to dinner and the invitation being accepted or not (Margutti et al. 2018). Linguistic research domains like register, politeness, or information density are significantly influenced by the socio-cultural set up of the interaction, as scholars like Hymes, Gumperz, Goffmann and Duranti (see above) have shown. The way language is used, with conventions in every dimension of social interaction, shapes the linguistic system described in grammars.

Even with a narrower focus on grammatical description, we maintain that the theoretical conception of language needs to be viewed as a multimodal system which should be recorded accordingly. To think that what we have called non-linguistic behaviour is not also governed by a set of rules is to have too narrow a perspective. We have not yet studied these elements to the same degree and in the same way, so we do not yet know all the rules that may exist. Theoretically, with a conception of language structure commensurate with multimodality, there is no reason to rule out the notion of rule-based multimodal behaviour, whereby the rules may be of a different nature and operate in different ways based on modality-specific differences in the use of sequentiality and simultaneity, categoricity and gradience.

The inability to see language as systematically incorporating different modes of meaning-making, and of incorporating a range of articulators, has also stood in the way of a linguistic treatment and analysis of sign languages as fully-fledged, natural languages. A concerted shift to a multimodal conception of language use would allow the full range of semiotic and multimodal resources, across both spoken and signed languages, to come into focus as constitutive of language. We agree with the need for theorizing language documentation (Himmelmann 2006) and argue for the need to conduct language documentation according to a theoretical conception of language as multimodal, negotiated and contingent. Before we turn to the implications for language documentation from a multimodal perspective (Section 6), we focus in the next section on what we stand to gain by studying multimodal language, including an overview of language-specific behaviour with respect to the coordination of multimodal cues.

5. What we stand to gain

When we document language according to the theoretical conception described in the previous section, we gain not only a record of communicative events as they occur in language use, in the face-to-face contexts that comprise the ecological niche of language. We also gain an understanding of language use, obtained through analyses of the diversity of language as a multimodal phenomenon and against the context of social interaction.

There are important testimonies to how the study of multimodal language gives us insight into the structure and conceptualization of language and of how multimodal utterances constitute language use and language structure. We offer examples of some of these important insights below. First, we take an empirical excursion into speech-accompanying manual gesture (Section 5.1). We focus on manual gestures because they show tight temporal and semantic integration with speech, providing evidence for gestures being part and parcel of language, and influencing and reflecting the relationship between language and cognition. Second, we provide some background on sign languages, focusing on different types of sign languages, signing practices and signing communities to help readers less familiar with sign languages understand what is at stake here (Section 5.2). Importantly, the late recognition of sign languages as fully-fledged, natural languages, the discrimination against deaf people and sign languages, and the challenges associated with creating records of sign language use, means that we have comparatively little knowledge and documented use of sign languages. We also stand to gain a fuller understanding of the diversity of language and cognition through a more widespread and systematic inclusion of sign languages in our documentation efforts.

5.1 Manual gestures

Seminal work by McNeill (1992) and Kendon (1986; 2004) sparked studies into the way speech and gesture are coordinated on a prosodic, semantic and pragmatic level, and the underlying cognitive basis for gesture and speech production. The temporal and functional coordination of speech and gesture has led to a general agreement that speech and gesture are two expressive modes of a single system. Gestures can be broadly grouped into referential and non-referential gestures (Kendon 2004), whereby referential gestures relate to the semantic content of an utterance and include representational or iconic gestures and pointing or deictic gestures, while non-referential gestures relate to the pragmatic content of an utterance. In this short overview, we provide examples that show that gestures are

integral to language and meaning generation and therefore should be recorded properly for a multipurpose record.

One context where the integral relationship between gesture, cognition, and language is very apparent is in the encoding of space and spatial relations. Haviland (1993), who described how spatial gestures are employed during storytelling among the Guugu Yimithirr (Pama-Nyungan, Queensland, Australia), provides an intriguing window into spatial cognition. Speakers of Guugu Yimithirr mainly employ an absolute frame of reference (cf. Levinson 2003), even in small-scale space, which means that they do not use relative terms like "right" and "left", and instead express location relations by means of cardinal directions. Haviland video-recorded a story about a boat that capsized and sank, which was again recorded two years later by Stephen Levinson. Analysing both recordings, he noticed that both times the (same) speaker used a speech-accompanying gesture showing how the boat rolled over sideways — however, the gesture was performed differently in the two instances. When the speaker was facing west, he gestured the rolling-over movement towards the front of his body, from east to west. Two years later, sitting facing north, he gestured the motion sideways from his right to the left side, and therefore still east to west. This shows that the absolute orientation prevalent in spatial reference in this speech community extends to the way memories of events are stored, a finding that contributed to the subsequent research focus on space within the Language and Cognition research group at the Max Planck Institute for Psycholinguistics.

Another important milestone in this research area was the work of David Wilkins. He also worked in Australia and described unique types of pointing gestures, used by speakers of Arrernte (Arnhem Land; Wilkins 2003). He found that formationally distinct gestures (e.g. in handshape, height of arm) are functionally distinct: referring, for example, to movement to a goal vs. the place itself; routes vs. directions; as well as referring (essentially categorically) to referents at different distances from the speaker. In addition, distinct handshapes are used to point to one vs. more than one referent. This is particularly interesting, given that there is no distinction in elements of speech in Arrernte between singular and plural. Speakers of Arrernte have to learn the categorical distinctions between different types of pointing gestures in the same way that other aspects of the language — e.g. the phonological and morphosyntactic constructions used in speech — are learned.

Mesh (2021) is a more recent example for an analysis and description of pointing gestures, describing the way in which modification of the height of a pointing gesture employed by speakers of the Quiahije variety of Eastern Chatino (Oaxaca, Mexico) is directly linked to the distance to the target, the "far-is-up strategy" (Mesh 2021: 1). As Mesh (2021) points out, while this seems to be a likely

candidate for a cross-linguistically and cross-culturally used gestural strategy (see also Wilkins 2003 for distance distinctions in Arrernte pointing), we do not have enough detailed descriptions of pointing gestures across languages and cultures to claim universality, or to understand better what competing motivations might lead to this strategy not being employed.

Finally, Ennever (2024) is a very recent example of a language description that includes speech-accompanying gesture. In this PhD thesis, Ennever provides a grammar of space of Kukatja (Pama-Nyungan, Western Australia). In addition to a description of all relevant aspects of the grammar (including spatial case marking, locative constructions and demonstratives), his thesis includes a chapter describing "aspects of pointing practices amongst the Kukatja" (Ennever 2024: 327). His findings reveal a complex system of pointing gestures employed for a variety of functions, including a particularly interesting "dual point" (Ennever 2024: 383), and a great importance of gesture in the encoding of complex motion events. His description clearly shows that gesture and speech are deeply integrated and need to be considered together in a grammatical description.

Pointing gestures are not restricted to spatial use. Floyd (2016: 35) describes his own auditory bias that at first made him miss temporal information in the multimodal utterances of the Nheengatú speakers (Tupi-Guarani, Brazil). Instead of providing precise temporal information in speech, a more generic spoken expression was combined with pointing to the sky, indicating the position of the sun at the intended time or across a time span. Floyd (2016: 48) convincingly argues for the linguistic nature of this multimodal encoding of time, and importantly also develops a "field elicitation method for composite utterances" that allows for "determining standards of form and productivity [...] for the visual modality."

The domain of negation and affirmation also received some attention recently (see e.g. an overview in Harrison 2024). Examples here are studies on gestures of the 'Away' family, a group of gestures that have in common a movement of the hands like sweeping, holding, brushing, or throwing something away. Bressem et al. (2017) present a study on Savosavo (Papuan, Solomon Islands), describing the various functions they identified for the 'sweeping away' (open hands are moved from the centre outwards like sweeping away crumbs from a surface) and 'holding away' (open hands are held up in front of the body or face with palms facing outward) gestures. There is a clear association of these gestures with negativity, but not necessarily negation, as they are more commonly employed when there is lexical or pragmatic (implicational) negation, and not so frequently found with explicit negation (Wegener & Bressem 2019). Based on this finding, Wegener and Bressem raised the question of functional load on gestures in relation to speech — their co-occurrence can emphasize explicit negation, and bring

lexical or implicit negation to the foreground. Inbar & Shor (2025) reported similar findings for Hebrew recently, which makes this pattern of use an interesting typological research topic in the future. Particularly relevant here is that the data for the Savosavo study was not collected with multimodal analysis in mind — most recordings made within the documentation project (funded by Volkswagen's DobeS project) were made with video showing most or all of the gesture space, enabling the later analysis of Savosavo gesture.

Gestures are also used with pragmatic functions, i.e. to indicate information structure or mark certain types of speech acts. Kendon (1995), for example, describes two different gesture forms ('mano-a-borsa' and 'mani giunte') used by speakers in Southern Italy to mark an utterance as a question or a request, or to indicate the topic-comment structure of utterances. Seyfeddinipur (2004) describes similar functions for the 'Pistol Hand' gesture (already documented in miniature paintings from the 15th century) used by speakers of Persian in the South of Iran and speakers of Azeri in the North of Iran. This gesture can accompany the comment part of an utterance (in contrast to the topic), and is also employed with directives. In this context it can even be used on its own, without speech (Seyfeddinipur 2004), underlining its substantial contribution to meaning generation and a potential conversion process to an emblematic gesture marking the speech act of an utterance.

Gestures that are not conventionalized can also provide important insights. Dingemanse (2013: 161) describes how Siwu speakers (Niger-Congo, Ghana) often combine ideophones and iconic gestures in "a single multimodal performative depictive act," illustrating the close connection and holistic multimodal generation of meaning. Moreover, he describes anecdotally how he was able to discern the difference in meaning between two ideophones after seeing the accompanying gesture depicting a circle with one ideophone versus depicting a sphere with the other. Studying Avatime (Niger-Congo, Ghana), Defina (2016) turned to spontaneous iconic gestures in order to answer the long-standing question of whether serial verb constructions (SVCs, verbal predicates of a clause consisting of two or more verb stems) reflect conceptualization of a complex event as a single event or as a combination of events. By looking at utterances containing SVCs and coding whether they are accompanied by one or more iconic gestures associated with the encoded event, she shows that single iconic gestures accompany SVCs, evidence for a conceptualization of a single, albeit complex, event. This is a prime example of the way that spontaneous, iconic gestures provide a "window onto thought" (McNeill 1992: 29), make conceptual structures accessible, and can be useful for both linguistic as well as psycholinguistic research questions.

5.2 Sign languages

When it comes to sign languages, the natural human languages used by deaf communities all over the world, the multimodality of language and the exceptionless need for video recording of language are obvious. The focus in this section is motivated by another question raised by the enterprise of language documentation, namely that of which languages to document. The primacy of speech and written text, as described above, has played a significant role in the exclusion and oversight of sign languages within linguistic documentation, description and analysis. This section provides some background on different types of sign languages and their ecologies of use, the nature of endangerment of sign languages, and their integration into documentary practice.

There are different kinds of classifications and typologies in use for sign languages (see Hou & de Vos 2022; Braithwaite 2019 for overviews). National, urban, or community sign languages refer to sign languages used primarily by deaf people in deaf communities (as opposed to hearing communities), e.g. Brazilian Sign Language (Libras). Their names typically reflect the nation-state or administrative region throughout which the sign language is used. Village, rural, or shared sign languages refer to sign languages used by deaf and hearing people (as reflected in the term "shared") in typically small rural communities with a comparatively high incidence of deafness. Examples include Kata Kolok in Bali or Adamorobe Sign Language in Ghana. Family sign languages refer to signing practices used in (extended) families with deaf people, whereby a larger signing community and sign language may emerge when different families come into contact with each other or co-reside in a space (Hou 2016). The term home signs (or home sign systems) overlaps with family sign language, but often refers to the use of signs by an individual to communicate with family members and friends. Nicaraguan Sign Language is an example of a community sign language that emerged through use across successive cohorts of signers, starting from a cohort of individual home signers, after they were brought together following the establishment of the first school for the deaf in Nicaragua (Kegl et al. 1999). Finally, secondary or alternate sign languages developed within hearing communities for use in specific contexts, for example, as part of rituals in periods of mourning, where speech is not allowed by cultural practice (Kendon 1984, 1987 on Warlpiri Sign Language). Alternate sign languages may also change status when deaf people enter or are born into the community (e.g. Yolngu Sign Language, Kendon 1988; Bauer 2014; Adone & Maypilama 2014).

The linguistic description of sign languages started in the second half of the 20th century, primarily with American Sign Language (ASL), moving to Western Europe, and then further across the globe. Recent decades have witnessed the

discovery and incipient description of many smaller sign languages with different ecologies and time-depths (see Le Guen et al. 2020). In addition, Zeshan & de Vos (2012) contains contributions about village sign languages around the globe, where rural shared community signing is the unifying element and time-depth may vary.

Ethnologue (Eberhard et al. 2025) currently lists 160 languages under the language family subgroup "Sign Language." Glottolog (Hammarström et al. 2025) groups sign languages under "Pseudo Family: Sign Language," listing a total of 227 sign languages. The grouping reflects the lack of systematic description and analysis of sign languages. Sign languages have their own genealogical and historical relationships, based on the histories of their communities: for example, American Sign Language (ASL) and French Sign Language (LSF) are related; American Sign Language (ASL) and British Sign Language (BSL) are not related. Compared to our knowledge of sign languages in North America and Western Europe, we still know little about sign languages existing or used in other parts of the world (Zeshan 2013; Braithwaite 2019). Documentary work can help to overcome this lack of knowledge and understanding.

Researchers working on the documentation of a spoken language may come into contact with or learn about a sign language being used somewhere nearby, or may learn about deaf individuals known to the community (Rarrick 2025). Expanding documentation efforts to include a sign language or sign system may be of great value and import, but it requires an understanding of and commitment to ethical and deaf-friendly research practice, and of the need for trust-building. This is true for any documentation project, of course, but is of particular importance in the context of sign languages and deaf people, where communities often experience discrimination, exclusion, and lack of access (see Singleton et al. 2015 for best-practice guidelines; see Rarrick 2025 and Braithwaite 2019 on encountering signing communities during linguistic fieldwork). In terms of the value of bridging between spoken and signed language documentation, Rarrick (2025) describes the important knowledge gain for linguistic analysis and insight into the multimodal nature of language use. Rarrick reports identifying a sign language, which she calls Sinasina Sign Language (SSSL), in use nearby her fieldsite during her documentation of Kere in PNG. She describes that gestures used by Kere speakers were subsequently identified as being lexical items from SSSL. The close relationship between the use of gestures by hearing members of the community and signs used in a geographically proximate sign language has been noted elsewhere (Coppola 2021; van Loon et al. 2014; Green 2021; Gawne & Cooperrider 2024). What seems special about the situation reported by Rarrick (2025) is that hearing speakers seem to be code-blending, i.e. using signs from the sign language simultaneously with speech (Emmorey et al. 2008), rather than the sign language

lexicalizing conventionalized or emblematic gestures from the surrounding hearing community.

As with spoken language documentation, sign language documentation is associated with language endangerment, and for similar reasons. These include the influence of a dominant language considered to be more prestigious; the association of a dominant language with better socio-economic opportunities; and the dispersal of a community due to migration to urban areas with better socio-economic opportunities. Hawai'i Sign Language (HSL) is an example of an endangered sign language which yielded to the use of the more dominant, prestigious, and institutionalized American Sign Language (ASL). The documentation of HSL suggests that it was an indigenous sign language in widespread use by the Hawai'i deaf community, but that its use declined when ASL was introduced into deaf schools in Hawai'i in the 1940s (Lambrecht et al. 2013; Woodward 2014). Mardin Sign Language (MarSL) is an example of a family sign language in rural south-eastern Turkey used by an extended family of forty individuals at the time of documentation, and with at least four generations of deaf individuals in the family (Zeshan & Dikyuva 2013). The language has become endangered primarily due to the dispersal of family members to urban areas like Istanbul and Izmir, in search of better employment opportunities. Deaf members of the family, especially younger individuals, no longer know MarSL, but rather Turkish Sign Language (TİD). Older deaf family members may still be bilingual in TİD and MarSL, which they use primarily with hearing MarSL signers in the family, all of whom are elderly.

The documentation of sign languages is of great importance for sign language linguistics and typology. We know comparatively little about the typological variation possible across sign languages (Perniss et al. 2007; Zeshan & Perniss 2008; Zeshan 2006 for comparative studies). In addition, investigations of the relationship between gesture and sign have important implications for our understanding of language and for language theory (cf. Kita & Emmorey 2023 for a recent review). For spoken languages, there are many domains that we cannot fully understand without looking at the use of gesture, or at least where we are getting much less than the full picture when we ignore gesture: the system of deixis with pointing; the use of positionals or posturals with iconic gestures; the use of negation with gestures and head movements; numerals and number gestures; spatial language and frame of reference systems and pointing. By the same token, we cannot fully understand these systems in human language, cognition and culture without looking at sign languages.

6. Implications for language documentation

Language documentation requires squaring theoretical concerns with practical realities. Himmelmann (2006: 4) underscores the need for theoretically grounded goals in language documentation to avoid "producing 'data graveyards,' i.e. large heaps of data with little to no use to anyone." Theory is relevant here to questions regarding the recording and processing of data and considerations of usability in the future. This theoretical grounding extends to the way that our discipline conceptualizes language: it influences the quality of the documentary collections, their use and usefulness, and the potential reach and impact of documentary collections now and in the long term. In short, providing a lasting, multipurpose record of language under a conception of language as multimodal and grounded in social structures and processes requires certain types of recording techniques and certain types of data. At the same time, every project will have its own specific goals and face its own particular demands and constraints, and thus pragmatic considerations will always play a role with respect to how theoretically grounded targets are achieved. These constraints include limited funding, tight timelines, technological hurdles, professional and personal responsibilities, community expectations, and, in some cases, graduation requirements.

In this section, we first outline implications of a multimodal conception of language for collecting the raw data, i.e. for video recording of language in use (Section 6.1). We then move to implications for deriving primary and structural data from the recordings, i.e. for providing transcriptions, translations and analyses of the language in use (Section 6.2). Finally, we outline implications for training, for linguists, in general, and field linguists, specifically (Section 6.3).

6.1 Implications for video recording

A conception of language as fundamentally multimodal and interactional necessarily requires video recordings of observable linguistic behaviour. For sign languages, the multimodal and interactional nature has always been clear, and recording of any part of sign language use is impossible without video recording. The technological advancements that have made the possibility of widespread video recording easy and accessible is an obvious boon to language documentation efforts. Schultze-Berndt (2015) links the change in the conceptualization of documentation practice — as consisting of a collection of audio-visual recordings — to these technological advancements. However, the technological possibility should not in itself be the reason for video recording, but instead should be theoretically motivated. Recording observable instances of language use within their social and cultural environments of use becomes part and parcel of description and analysis

of language, and thus our understanding of language. This theoretical imperative for video recording puts some demands on the implementation. Recording needs to be done in a way that allows primary and structural data to be usefully derived from it.

Unlike ethnographic film, which uses aesthetic means for storytelling, documentary recording should be as neutral and rich as possible. This influences the choice of equipment as well as the framing of an event that is being recorded. For example, SLR (single-lens reflex) cameras with depth of field blur the background, and thus blur others who may also be present during the recording or blur the location of the recording, both of which may influence how the recorded speaker talks (e.g. avoiding the choice of certain topics or words when children are present). As such, this aesthetic recording technique defies the purpose of attempting to record as much information as possible given the technical and practical restrictions (Seyfeddinipur & Rau 2020). This is not to say that these recordings have no place in language documentation; they may, for example, be useful for the creation of outreach material (Dimmendaal 2010). For documentation, the framing of scenes should include the whole gesture/sign space, meaning the space around the body in which gestures/signs are produced, including the reach of pointing with an outstretched arm. In addition, the interactional partners, i.e. the people being talked to or with, should be included in the recording, so that it is possible to understand choices related to, for example, register, politeness level, information density, turn taking patterns, feedback and other coordination techniques.

Making recordings that capture language use in its natural environment also means balancing between theoretical desiderata and practical constraints. The placement of cameras and microphones is crucial. Equipment should be positioned as neutrally as possible to interfere with the situation as little as possible, but also as strategically as possible to capture as much of the environment and participants as possible. For a long time, recording video in the field was challenging due to the high cost of cameras and storage, lack of electricity in the field, and processing capacities of laptops. In recent years, with the advent of affordable action cameras, it has become much easier to record different types of communicative events; walking through a landscape while speakers describe the environment, while pointing, for example, to different plants, animals, or spaces (e.g. Burenhult 2020). With 360° cameras, which afford a full circle field of view, it is now possible to record speakers/signers sitting in circular configurations (González Pérez, submitted). Two exemplary collections, Wulff (2023) and González Pérez & Conrad (2023), in which the documenters use action cameras for capturing interactions in their natural contexts in different spatial configurations can be found in the Endangered Languages Archive (ELAR). Small

lightweight vlogger cameras now often include built-in or compatible gimbals that allow for mechanical stabilization, making footage much smoother and thus useful for situations where hand-held camera use is necessary, as in market situations. The availability of multiple cameras or different types of cameras allows for more sophisticated, multi-perspective recording of language use situations. However, there are obvious constraints on such availability, ranging from limited funding for equipment to limited processing capabilities and lack of electricity. At the same time, technological advances reinforce the feasibility of the approach suggested here. While even a decade ago the use of video posed considerable logistical and financial challenges, today's recording equipment, including mobile phones, is compact, lightweight, comparatively affordable and highly versatile. Similarly, the decreasing cost and increasing capacity of digital storage make it possible to preserve large-scale, high-fidelity audio-visual data with relative ease.

Coming back to theoretical considerations, a central principle in this approach is the commitment to documenting language use in its fullest form, based on the idea that interactional dynamics and contextual factors are integral to understanding linguistic practice, even when particular phenomena — such as gestures, prosodic features, turn-taking, or repair sequences — are not the explicit focus of a given project. Including a proper recording of the interaction facilitates a systematic investigation of this complex interaction, and can allow the documentary multipurpose record to serve many potential analyses not possible or even envisioned at the time of the documentation project. By capturing a broad and authentic record of language use as it is practiced in everyday contexts, we provide future researchers and community members with a multipurpose record that supports a wide range of inquiries. Importantly, this means that the raw recorded data must be accompanied by detailed metadata on the setting, participants and context. In addition, archiving is necessary to ensure that the data is available and usable at a later date.

6.2 Implications for transcription and analysis

The provision of primary data — transcript and translation — from the raw data requires linguistic and native speaker/signer expertise, and every level of annotation depends on the particular focus of the documenter. A community-driven project documenting traditional narratives may require a close transcription and in the second step a transliteration and translation into the major contact language. A project documenting an undescribed language may start with a close IPA-based transcription and translation into English, with an orthographic transcription and glossed text at a later point in time, once the phonological system has been understood. The annotation of gesture or sign requires expertise on the

formational transcription of manual and non-manual actions; annotation of turn-taking requires expertise in the transcription of turns and other interactional features, and so on.

We align with Himmelmann (2006; 2011; 2018) in taking documentary linguistics to comprise raw data and primary data, and the relationships between them. We recognize that it will not be possible for individual researchers or community members to provide full transcripts and translations of (all) video-recorded linguistic behaviour (of language in use). As such, descriptive linguistics — deriving analyses and grammatical descriptions from the transcripts and translations — will not be possible across the board in individual documentation projects. We come back to the notion of concentric circles: if individuals are themselves experts in a domain, or have sufficient knowledge in a domain, e.g. in the recognition of gestures, some primary data can be provided with respect to the transcription of multimodal cues. Similarly, some annotation and analysis can be provided. This can help to offer a framework for transcription, and may be accompanied by suitable training material. Again, the implications for documentation are to try to maximize the potential for later multimodal transcription and analysis. Anything that may help make (grammatical) descriptions better or more precise on any level would be valuable. This could include notes or minimal marking in the transcript files, e.g. noting that pointing is used regularly in reference ambiguity resolution and marking the occurrence of (potential) pointing gestures when they are observed. Importantly, the practical realities of documentation may place a great deal of pressure on those undertaking the work, whether academics or community members. Placing even further demands on documenters is neither practical nor sustainable. In addition, the time and effort involved in creating a transcribed and annotated audio-visual corpus is still not highly rewarded in academia's merit system. This can be a problem especially for early-career scholars.

6.3 Implications for training

To embark on language documentation projects according to a framework of language as fundamentally interactive and multimodal requires training that conveys such a conception of language. We address three main issues with implications for training. The first concerns what we think of when we think of language. As we have stressed throughout this chapter, language should be thought of not just as speech or text and primarily monologic in nature, but as multimodal, situated and interactional. To entrench this, including the need for video recording of data, curricula for language documentation theory and methods should incorporate key insights from ethnomethodology, conversation analysis, anthro-

pology, linguistic anthropology and sociology relating to essential theoretical and methodological tools for understanding language use. For example, ethnographic observation is crucial for grasping interactional structures and processes and for ensuring that recording situations align with the cultural norms governing language use. In addition, training must ensure that recordings are created with an understanding of why and how individuals communicate in a given situation and environment.

The second issue concerns what languages we think of when we think about languages: that is, not just spoken languages, but also sign languages. Knowledge of sign languages — including knowledge of different types of sign languages and different ecologies of use, and of the nature of endangerment of sign languages — must be integrated into training. In thinking of the landscape of languages and the fabric of linguistic diversity, sign languages should not appear as an afterthought or exception, but as integral to the constellation of languages across the world. Our scope and definition of language documentation should thus necessarily extend to sign languages. In addition, knowledge of sign languages must be embedded into training in documentary theory and practice for the practical reason that small-scale sign languages (or sign systems) are often encountered in the process of documentation projects on spoken languages.

The third point concerns our notion of linguistic description. A reference grammar, for example, would benefit from expanding grammatical descriptions based on structures gleaned from speech to incorporate integral and concomitant multimodal components. Training should be provided in the already established baseline findings coming from gesture research. Deixis and negation, for example, are two areas where gestural components can be integral. The co-occurrence of deictic elements in speech and gesture is ubiquitous (and in some cases may even be obligatory, cf. Wilkins 2003; Levinson et al. 2018; Sherzer 1973). A head shake may function as a negation expression in its own right, fully replacing the verbal component (Kendon 2002). However, in practice, while standard reference grammars may refer to the use of gestures, they are typically not described in detail or integrated into analysis. If they were, the analysis of deixis (including reference resolution and the notion of grammatical ambiguity with respect to pronoun reference) would likely look different. From the perspective of language documentation as we have argued here, the structural possibilities of a language are enriched with the multimodal aspects of the language.

7. Conclusion

The multimodality of language and the dynamics of sociocultural practices of language use are implicit — and in many ways also explicit — in the call to language documentation to provide a record of language use, and in the fundamental understanding of language as a social practice and cognitive faculty. We recommend a close reading of Himmelmann (1998), (2006) and (2012), where the foundations are clearly laid out. To understand the diversity of human language, cognition and culture and the human capacity for language, we must document this diversity and the negotiation and creation of meaning in interaction. In short, we must take language use seriously in language documentation so the record can be multipurpose and ultimately serve as the primary resource for detailed linguistic analysis of structural and social aspects of language use, as well as providing a crucial foundation for language maintenance and revitalization initiatives.

Acknowledgements

The authors would like to thank the editors and anonymous reviewers for their valuable and constructive comments, which greatly contributed to improving the final version of this paper.

References

Adone, Marie Carla D., and Elaine Maypilama. 2014. *A Grammar Sketch of Yolngu Sign Language*. München: Lincom Europe.

Albert, Saul, and Jan P. de Ruiter. 2018. "Repair: The Interface between Interaction and Cognition." *Topics in Cognitive Science* 10 (2): 279–313.

Bangerter, Adrian, and Herbert H. Clark. 2003. "Navigating Joint Projects with Dialogue." *Cognitive Science* 27 (2): 195–225.

Bauer, Anastasia. 2014. *The Use of Signing Space in a Shared Sign Language of Australia*. Berlin, Boston: De Gruyter Mouton.

Bavelas, Janet Beavin. 2022. *Face-to-Face Dialogue: Theory, Research, and Applications*. Oxford: Oxford University Press.

Bolden, Galina B. 2011. "On the Organization of Repair in Multiperson Conversation: The Case of 'Other'-Selection in Other-Initiated Repair Sequences." *Research on Language and Social Interaction* 44 (3): 237–262.

Bowern, Claire. 2015. *Linguistic Fieldwork: A Practical Guide*. Basingstoke: Palgrave Macmillan.

Braithwaite, Ben. 2019. "Sign Language Endangerment and Linguistic Diversity." *Language* 95 (1): e161–e187.

Bressem, Jana, Nicole Stein, and Claudia Wegener. 2017. "Multimodal Language Use in Savosavo: Refusing, Excluding and Negating with Speech and Gesture." *Pragmatics* 27 (2): 173–206.

Burenhult, Niclas. 2020. "Domain-driven Documentation: The Case of Landscape." In *Approaches to Interdisciplinary Research in Language Documentation, Language Documentation and Conservation*, ed. by Susan Penfield, 9–22. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24943

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Clark, Herbert H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Clark, Herbert H., and Jean E. Fox Tree. 2002. "Using 'uh' and 'um' in spontaneous speaking." *Cognition* 84 (1): 73–111.

Clark, Herbert H., and Richard J. Gerrig. 1990. "Quotations as Demonstrations." *Language* 66 (4): 764–805.

Coppola, Marie. 2021. "Sociolinguistic Sketch: Nicaraguan Sign Language and Homesign Systems in Nicaragua." In *Emerging Sign Languages of the Americas*, ed. by Olivier Le Guen, Josefina Safar, and Marie Coppola, 439–450. Berlin, Boston: De Gruyter Mouton.

Defina, Rebecca. 2016. "Do Serial Verb Constructions Describe a Single Event? A Study of Co-speech Gestures in Avatime." *Language* 92 (4): 890–910.

Dimmendaal, Gerrit J. 2010. "Language Description and 'The New Paradigm': What Linguists May Learn from Ethnocinematographers." *Language Documentation and Conservation* 4: 152–158. http://hdl.handle.net/10125/4475

Dingemanse, Mark. 2013. "Ideophones and Gesture in Everyday Speech." *Gesture* 13 (2): 143–165.

Dingemanse, Mark, and Nick J. Enfield. 2015. "Other-initiated Repair Across Languages: Towards a Typology of Conversational Structures." *Open Linguistics* 1 (1): 98–118.

Dingemanse, Mark, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and Nick J. Enfield. 2015. "Universal Principles in the Repair of Communication Problems." *PLOS ONE* 10 (9): e0136100.

Duerr, Michael. 2013. "Die Erforschung der Sprachen Kamčatkas [Research into the Kamčatkas languages]." In *Reisen an den Rand des Russischen Reiches: Die wissenschaftliche Erschließung der nordpazifischen Küstengebiete im 18. und 19. Jahrhundert* [*Journeys to the edge of the Russian empire: The scientific exploration of the North Pacific coastal regions in the 18th and 19th centuries*], ed. by Erich Kasten, 287–296. Fürstenberg/Havel: Kulturstiftung Sibirien.

Duranti, Alessandro. 1997. *Linguistic Anthropology*. Cambridge: Cambridge University Press.

Duranti, Alessandro. 2005. "History of Linguistic Anthropology." In *Linguistic Anthropology: Encyclopedia of Life Support Systems (EOLSS)*, ed. by Anita Sujodzic, 263–278. Paris: EOLSS Publishers.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds). 2025. *Ethnologue: Languages of the World*, 28th Edition. Dallas, Texas: SIL International. Online version: https://www.ethnologue.com/

Emmorey, Karen, Helsa B. Borinstein, Robin Thompson, and Tamar H. Gollan. 2008. "Bimodal Bilingualism." *Bilingualism: Language and Cognition* 11 (1): 43–61.

Ennever, Thomas Blake. 2024. Topics in a Grammar of Space of Kukatja. PhD Thesis, Monash University.

Evans, Nicholas. 2008. "Review of Essentials of Language Documentation. 2006." *Language Documentation and Conservation* 2 (2): 340–350. http://hdl.handle.net/10125/

Floyd, Simeon. 2016. "Modally Hybrid Grammar? Celestial Pointing for Time-of-day Reference in Nheengatú." *Language* 92 (1): 31–64.

Garfinkel, Harold. 1967. *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Gawne, Lauren, and Kensy Cooperrider. 2024. "Emblems: Meaning at the Interface of Language and Gesture." *Glossa: A Journal of General Linguistics* 9 (1).

Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel. 2006. *Essentials of Language Documentation*. Berlin, New York: De Gruyter Mouton.

Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. New York: Anchor.

González Pérez, Manuel David. Submitted. "Widening the View: Immersive 360° Filming in Language Documentation." *Language Documentation and Description*.

González Pérez, Manuel David, and Agnes Conrad. 2023. *Language, Environment and Social Cognition in Southeastern Yunnan: A Holistic Documentation of Multimodal Communication amongst the Phola*. Endangered Languages Archive. http://hdl.handle.net/2196/5830a66e-1384-484a-91cb-52160a0cdbcd

Goodwin, Charles. 1986. "Gestures as a Resource for the Organization of Mutual Orientation." *Semiotica* 62 (1–2): 29–49.

Goodwin, Charles. 2014. "The Intelligibility of Gesture Within a Framework of Co-operative Action." In *From Gesture in Conversation to Visible Action as Utterance: Essays in Honor of Adam Kendon*, ed. by Mandana Seyfeddinipur, and Marianne Gullberg, 199–216. Amsterdam: John Benjamins.

Goodwin, Marjorie Harness, and Charles Goodwin. 1986. "Gesture and Coparticipation in the Activity of Searching for a Word." *Semiotica* 62 (1–2): 51–75.

Green, Jennifer. 2021. "Mothers, Mountain Devils, and Pointing to Eternity: The 'Horns' Handshape in Australian Indigenous Sign Languages." *Sign Language Studies* 22 (1): 5–41.

Gumperz, John J., and Dell Hymes (eds). 1964. "The Ethnography of Communication." Special Issue, *American Anthropologist* 66 (6.2). www.jstor.org/stable/i227427

Hale, Kenneth, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. 1992. "Endangered Languages." *Language* 68 (1): 1–42.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. *Glottolog 5.2*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Hanks, William F. 2005. "Explorations in the Deictic Field." *Current Anthropology* 46 (2): 191–220.

Harrison, Simon. 2024. "On Grammar-Gesture Relations: Gestures Associated with Negation." In *The Cambridge Handbook of Gesture Studies*, ed. by Alan Cienki, 446–474. Cambridge: Cambridge University Press.

Haviland, John B. 1993. "Anchoring, Iconicity, and Orientation in Guugu Yimithirr Pointing Gestures." *Journal of Linguistic Anthropology* 3 (1): 3–45.

Henley, Paul. 2020. *Beyond Observation: A History of Authorship in Ethnographic Film*. Manchester: Manchester University Press.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2006. "Language Documentation: What Is It and What Is It Good For?" In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin, New York: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2011. On Linguistic Uses of Language Documentations. In *Selected Papers from the International Conference on Language Documentation and Tradition (7–9 November 2008)*, ed. by Carol Everhard, Elizabeth Mela-Athanasopoulou, and Youli Theodosiadou, 37–51. Thessaloniki: Aristotle University of Thessaloniki, University Studio Press.

Himmelmann, Nikolaus P. 2012. "Linguistic Data Types and the Interface between Language Documentation and Description." *Language Documentation and Conservation* 6: 187–207. http://hdl.handle.net/10125/4503

Himmelmann, Nikolaus P. 2018. "Meeting the Transcription Challenge." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 33–40. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24806

Hinton, Leanne, Leena Huss, and Gerald Roche (eds.). 2018. *The Routledge Handbook of Language Revitalization*. New York: Routledge.

Hou, Lynn Y.-S. 2016. "Making Hands": Family Sign Languages in the San Juan Quiahije Community. PhD Thesis, The University of Texas at Austin. http://hdl.handle.net/2152/46879

Hou, Lynn, and Connie de Vos. 2022. "Classifications and Typologies: Labeling Sign Languages and Signing Communities." *Journal of Sociolinguistics* 26 (1): 118–125.

Humphries, Tom. 1977. Communicating Across Cultures (Deaf-Hearing) and Language Learning. PhD Thesis, Union Institute and University. ProQuest (DP10817).

Hymes, Dell. 1962. "The Ethnography of Speaking." In *Anthropology and Human Behavior*, ed. by Thomas Gladwin, and William C. Sturtevant, 13–53. Washington, D.C.: Anthropology Society of Washington.

Hymes, Dell. 1964. "Toward Ethnographies of Communication." *American Anthropologist* 66 (6): 1–34. https://www.jstor.org/stable/668159

Inbar, Anna, and Leon Shor. 2025. "The Depiction of Negativity: The Case of the Brushing Hands Gesture Used by Hebrew Speakers." In *Perspectives on Negation: Views from across the Language Sciences*, ed. by Frances Blanchette, and Cynthia Lukyanenko, 21–56. Berlin, Boston: De Gruyter Mouton.

Jefferson, Gail. 2004. "Glossary of Transcript Symbols with an Introduction." In *Conversation Analysis: Studies from the First Generation*, ed. by Gene H. Lerner, 13–31. Amsterdam: John Benjamins.

Kegl, Judy, Ann Senghas, and Marie Coppola. 1999. "Creation through Contact: Sign Language Emergence and Sign Language Change in Nicaragua." In *Language Creation and Language Change: Creolization, Diachrony, and Development*, ed. by Michel DeGraff, 179–237. Cambridge, MA: MIT Press.

Kendon, Adam. 1967. "Some Functions of Gaze-direction in Social Interaction." *Acta Psychologica* 26: 22–63.

Kendon, Adam. 1979. "Some Emerging Features of Face-to-Face Interaction Studies." *Sign Language Studies* 22: 7–22.

Kendon, Adam. 1983. "The Study of Gesture: Some Remarks on its History." In *Semiotics 1981*, ed. by John N. Deely, and Margot D. Lenhart, 153–164. Boston, MA: Springer.

Kendon, Adam. 1984. "Knowledge of Sign Language in an Australian Aboriginal Community." *Journal of Anthropological Research* 40 (4): 556–576.

Kendon, Adam. 1986. "Some Reasons for Studying Gesture." *Semiotica* 62 (1–2): 3–28.

Kendon, Adam. 1987. "Speaking and Signing Simultaneously in Warlpiri Sign Language Users." *Multilingua* 6 (1), 25–68.

Kendon, Adam. 1988. *Sign Languages of Aboriginal Australia: Cultural, Semiotic, and Communicative Perspectives*. Cambridge: Cambridge University Press.

Kendon, Adam. 1995. "Gestures as Illocutionary and Discourse Structure Markers in Southern Italian Conversations." *Journal of Pragmatics* 23 (3): 247–279.

Kendon, Adam. 2002. "Some Uses of the Head Shake." *Gesture* 2 (2): 147–182.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kita, Sotaro, and Karen Emmorey. 2023. "Gesture Links Language and Cognition for Spoken and Signed Languages." *Nature Reviews Psychology* 2: 407–420.

Krauss, Michael. 1992. "The World's Languages in Crisis." *Language* 68 (1): 4–10.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lambrecht, Linda, Barbara Earth, and James Woodward. 2013. "History and Documentation of Hawai'i Sign Language: First Report." Presentation, 3rd International Conference on Language Documentation and Conservation, University of Hawai'i, Honolulu, 28 February–7 March 2013. http://hdl.handle.net/10125/26133

LaPolla, Randy J. 2020. "Forward to the Past: Modernizing Linguistic Typology by Returning to its Roots." *Asian Languages and Linguistics* 1 (1): 147–167.

Le Guen, Olivier, Josefina Safar, and Marie Coppola (eds). 2020. *Emerging Sign Languages of the Americas*. Berlin, Boston: De Gruyter Mouton.

Levinson, Stephen C. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.

Levinson, Stephen C., Sarah Cutfield, Michael J. Dunn, Nick J. Enfield, and Sérgio Meira (eds). 2018. *Demonstratives in Cross-Linguistic Perspective*. Cambridge: Cambridge University Press.

Linell, Per. 2004. *The Written Language Bias in Linguistics: Its Nature, Origins and Transformations*. London: Routledge.

MacDougall, David. 1998. *Transcultural Cinema*. Princeton: Princeton University Press.

Margutti, Piera, Liisa Tainio, Paul Drew, and Véronique Traverso. 2018. "Invitations and Responses across Different Languages: Observations on the Feasibility and Relevance of a Cross-linguistic Comparative Perspective on the Study of Actions." *Journal of Pragmatics* 125: 52–61.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal About Thought*. Chicago: The University of Chicago Press.

Meakins, Felicity, Jennifer Green, and Myfany Turpin. 2018. *Understanding Linguistic Fieldwork*. London: Routledge.

Mesh, Kate. 2021. "It's as Far as the Arm Can Raise: Pointing Height Marks Target Distance among the San Juan Quiahije Chatino." *Lingua* 259: 103099.

Mondada, Lorenza. 2008. "Using Video for a Sequential and Multimodal Analysis of Social Interaction: Videotaping Institutional Telephone Calls." *Forum: Qualitative Social Research* 9 (3): 39.

Ochs, Elinor. 1988. *Culture and Language Development: Language Acquisition and Language Socialization in a Samoan Village*. Cambridge: Cambridge University Press.

Ong, Walter J. 1982. *Orality and Literacy: The Technologizing of the Word*. London, New York: Routledge.

Pakendorf, Brigitte, and Françoise Rose (eds). 2025. *Fillers: Hesitatives and Placeholders*. Berlin: Language Science Press.

Parry, Ruth. 2010. "Video-based Conversation Analysis." In *The SAGE Handbook of Qualitative Methods in Health Research*, ed. by Ivy Bourgeault, Robert Dingwall, and Raymond De Vries, 373–396. London: Sage.

Perniss, Pamela. 2018. "Why We Should Study Multimodal Language." *Frontiers in Psychology* 9: 1109.

Perniss, Pamela, Roland Pfau, and Markus Steinbach (eds). 2007. *Visible Variation: Comparative Studies on Sign Language Structure*. Berlin, New York: De Gruyter Mouton.

Rarrick, Samantha. 2025. "Bridging Signed Language Documentation and Spoken Language Documentation." *Language Documentation and Conservation* 19: 24–39. https://hdl.handle.net/10125/74809

Sacks, Harvey. 1992. *Lectures on Conversation*, Vol. 1 and 2, ed. by Gail Jefferson. Oxford: Blackwell Publishing.

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. "A Simplest Systematics for the Organization of Turn-taking for Conversations." *Language* 50 (4.1): 696–735.

Saussure, Ferdinand de. 1916. *Cours de linguistique générale*. Lausanne, Paris: Payot.

Schegloff, Emanuel A. 1982. "Discourse as an Interactional Achievement: Some Uses of 'uh huh' and Other Things that Come between Sentences." In *Analyzing Discourse: Text and Talk*, ed. by Deborah Tannen, 71–93. Washington, D.C.: Georgetown University Press.

Schegloff, Emanuel A. 1998. "Body Torque." *Social Research* 65 (3): 535–596. https://www.jstor.org/stable/40971262

Schegloff, Emanuel A. 2009. "A Practice for (Re-)Exiting a Sequence: And/But/So + Uh(m) + Silence." In *Language in Life, and a Life in Language: Jacob Mey — A Festschrift*, ed. by Bruce Fraser, and Ken Turner, 365–374. Bingly: Emerald.

Schegloff, Emanuel A., Gail Jefferson, and Harvey Sacks. 1977. "The Preference for Self-Correction in the Organization of Repair in Conversation." *Language* 53 (2): 361–382.

Schultze-Berndt, Eva. 2015. "Language Documentation." In *Syntax: Theory and Analysis: An International Handbook*, Vol. 3, ed. by Tibor Kiss, and Artemis Alexiadou, 2063–2094. Berlin, Boston: De Gruyter Mouton.

Seyfeddinipur, Mandana. 2004. "Meta-discursive Gestures from Iran: Some Uses of the 'Pistol Hand'." In *The Semantics and Pragmatics of Everyday Gestures*, ed. by Cornelia Müller, and Roland Posner, 205–216. Berlin: Weidler.

Seyfeddinipur, Mandana. 2012. "Reasons for Documenting Gestures and Suggestions for How to Go About It." In *The Oxford Handbook of Linguistic Fieldwork*, ed. by Nick Thieberger, 147–165. Oxford: Oxford University Press.

Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L. Epps, Vera Ferreira, Ana Vilacy Galucio, Birgit Hellwig, Oliver Hinte, Gary Holton, Dagmar Jung, Irmgarda Kasinskaite Buddeberg, Manfred Krifka, Susan Kung, Miyuki Monroig, Ayu'nwi Ngwabe Neba, Sebastian Nordhoff, Brigitte Pakendorf, Kilu von Prince, Felix Rau, Keren Rice, Michael Riessler, Vera Szoelloesi-Brenig, Nick Thieberger, Paul Trilsbeek, Hein van der Voort, and Tony Woodbury. 2019. "Public Access to Research Data in Language Documentation: Challenges and Possible Strategies." *Language Documentation and Conservation* 13: 545–563. http://hdl.handle.net/10125/24901

Seyfeddinipur, Mandana, and Felix Rau. 2020. "Keeping it Real: Video Data in Language Documentation and Language Archiving." *Language Documentation and Conservation* 14: 503–519. http://hdl.handle.net/10125/24965

Sherzer, Joel. 1973. "Verbal and Nonverbal Deixis: The Pointed Lip Gesture among the San Blas Cuna." *Language in Society* 2 (1): 117–131.

Sherzer, Joel. 1983. *Kuna Ways of Speaking: An Ethnographic Perspective*. Austin: The University of Texas Press.

Singleton, Jenny L., Amber J. Martin, and Gary Morgan. 2015. "Ethics, Deaf-friendly Research, and Good Practice When Studying Sign Languages." In *Research Methods in Sign Language Studies: A Practical Guide*, ed. by Eleni Orfanidou, Bencie Woll, and Gary Morgan, 5–20. Oxford: Wiley-Blackwell?.

Streeck, Jürgen. 2009. *Gesturecraft: The Manu-facture of Meaning*. Amsterdam: John Benjamins.

Swerts, Marc. 1998. "Filled pauses as markers of discourse structure." *Journal of Pragmatics* 30 (4): 485–496.

van Loon, Ester, Roland Pfau, and Markus Steinbach. 2014. "The Grammaticalization of Gestures in Sign Languages." In *Body — Language — Communication: An International Handbook on Multimodality in Human Interaction*, Vol. 2, ed. by Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill, and Jana Bressem, 2133–2149. Berlin, Boston: De Gruyter Mouton.

Vigliocco, Gabriella, Pamela Perniss, and David Vinson. 2014. "Language as a Multimodal Phenomenon: Implications for Language Learning, Processing and Evolution." *Philosophical Transactions of the Royal Society B* 369 (1651): 20130292.

Wegener, Claudia, and Jana Bressem. 2019. "Sharing the Load — The Interplay of Verbal and Gestural Negation in Savosavo." Presentation, LingCologne 2019, University of Cologne, Cologne, 6–7 June 2019.

Wilkins, David. 2003. "Why Pointing with the Index Finger is Not a Universal (in Sociocultural and Semiotic Terms)." In *Pointing: Where Language, Culture and Cognition Meet*, ed. by Sotaro Kita, 171–215. Mahwah, NJ: Laurence Erlbaum.

Woodbury, Anthony C. 2003. "Defining Documentary Linguistics." In *Language Documentation and Description*, Vol. 1, ed. by Peter K. Austin, 35–51. London: SOAS.

Woodbury, Anthony C. 2011. "Language Documentation." In *The Cambridge Handbook of Endangered Languages*, ed. by Peter K. Austin, and Julia Sallabank, 159–186. Cambridge: Cambridge University Press.

Woodward, James. 2014. *Documentation of Hawaii Sign Language: Building the Foundation for Documentation, Conservation, and Revitalization of Endangered Pacific Island Sign Languages*. Endangered Languages Archive. http://hdl.handle.net/2196/00-0000-0000-0007-61D1-6

Wulff, Mareike. 2023. *Documenting Lhokpu Language in Bhutan and the Endangered Yarn Production Practices*. Endangered Languages Archive. http://hdl.handle.net/2196/f65ac9a9-f1ba-44bd-a310-6df3daaeaf04

Zeshan, Ulrike (ed). 2006. *Interrogative and Negative Constructions in Sign Language*. Nijmegen: Ishara Press.

Zeshan, Ulrike. 2013. "Sign Languages." In *The World Atlas of Language Structures Online*, v2020.4, ed. by Matthew S. Dryer, and Martin Haspelmath. Zenodo. (available online at http://wals.info/chapter/s9).

Zeshan, Ulrike, and Pamela Perniss (eds). 2008. *Possessive and Existential Constructions in Sign Languages*. Nijmegen: Ishara Press.

Zeshan, Ulrike, and Connie de Vos (eds). 2012. *Sign Languages in Village Communities: Anthropological and Linguistic Insights*. Berlin, Boston: De Gruyter Mouton.

Zeshan, Ulrike, and Hasan Dikyuva. 2013. "Documentation of Endangered Sign Languages: The Case of Mardin Sign Language." In *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*, ed. by Mari C. Jones, and Sarah Ogilvie, 29–41. Cambridge: Cambridge University Press.

CHAPTER 9

The multimodal turn in corpus building

Lessons from language documentation and sign language corpora

Anastasia Bauer, Sonja Gipper,
Tobias-Alexander Herrmann, Jana Hosemann
& Simona Sbranna
University of Cologne

This chapter argues for a decisive shift in corpus-building practices toward a *Multimodal Turn*. For decades, smaller-scale spoken and signed languages were less available for linguistic corpus research due to the scarcity of corpora, but the situation has changed: language documentation now provides rich multimodal resources, and many sign languages are recorded in high-definition video, offering unique insights into communicative practices. In contrast, corpora of majority languages often remain limited in video data. We contend that they can draw inspiration from the practices developed in language documentation and sign language research to close this gap. Only by systematically integrating multimodal interaction into corpus design can linguistic research capture the full complexity of human communication.

Keywords: multimodality, corpus building, language documentation, sign languages, multimodal feedback

1. Introduction

One of the central aims of linguistics is to account for the rich heterogeneity of human language. Looking both at the history of the field and at contemporary research, we observe a growing shift towards exploring the interactional-pragmatic dimensions of language use. Increasingly, linguists recognize that the human language capacity is deeply rooted in social interaction (Levinson 2006; Holler & Levinson 2019; Macuch Silva et al. 2020; Hamilton & Holler 2023). Crucially, this recognition goes hand in hand with the insight that language is inherently multimodal (Perniss 2018; Özyürek 2021; Kendrick et al. 2023), and that face-to-face,

https://doi.org/10.1075/slcs.240.09bau
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

i.e., co-present (Ameka & Terkourafi 2019) communication invariably draws on a rich repertoire of interwoven multi-semiotic practices that create meaning together (Enfield 2009; Ferrara & Hodge 2018).

Language corpora constitute one of the central resources in linguistics, providing collections of language data that enable the systematic study of patterns of language use, the testing of hypotheses, and the investigation of variation and change in a data-driven manner. The recognition of corpora as a fundamental resource for linguistic inquiry dates back several decades (Chafe 1992). Yet, despite growing attention to multimodality, there remains a notable scarcity of truly multimodal corpora for highly resourced, well-described spoken majority languages such as English. Instead, corpora of majority languages are often compiled in large parts from written language samples or (semi-)scripted spoken language such as TV shows (see Section 3.3). This gap is striking given that corpora are uniquely positioned to present language as it unfolds in its natural habitat.

In this chapter, we argue that majority spoken language corpora can benefit profoundly from the approaches developed in language documentation and sign language corpus building. Such approaches offer inspirations for building truly multimodal, multipurpose corpora — a vision in line with the pioneering ideas of Nikolaus Himmelmann (1998) and his call for richer, more holistic documentation of language use. Such resources would not only align with the evolving priorities of linguistic research but also provide the robust empirical foundation necessary for understanding language as a fundamentally multimodal phenomenon.

We start by discussing the importance of multimodality in language research in Section 2. The three broad types of corpora considered in this chapter — language documentation corpora, sign language corpora, and majority language corpora — are introduced in Section 3. In Section 4, we discuss how majority language corpus building can benefit from the approaches developed in language documentation and sign language research. We conclude our discussion in Section 5.

2. Multimodality and why it matters for language research

Scholars across the language sciences and related disciplines increasingly recognize that communication is inherently multimodal (Bavelas et al. 1995; Kendon 2004; Vigliocco et al. 2014; Abner et al. 2015; Mondada 2016; Keevallik 2018; Perniss 2018; Hagoort & Özyürek 2025; Sandler 2024), a shift that has been characterized as a *Multimodal Turn* (Ravelli 2024). Consequently, there is a growing demand for multimodal corpora that far outpaces their current availability (Allwood 2008).

Although the number of investigations into multimodal communication is steadily growing, the notion of multimodality remains broad and is interpreted in diverse ways (Partan & Marler 1999; Sandler 2022). To provide clarity, we first outline our understanding of multimodality in Section 2.1. In Section 2.2, we present a case study from our own research that illustrates the important role of bodily signals such as head movements in everyday human co-present interaction. This case study underscores the need to move beyond unimodal perspectives and to place greater emphasis on multimodal phenomena in language research.

2.1 What is multimodality?

The term multimodality has been employed in linguistics and related disciplines, but often with different emphases and definitions. There are two broad perspectives: the first emphasizes language and communication as inherently multimodal phenomena, while the second characterizes the kinds of data collected and analysed.

On the first view, language is recognized as spanning different channels of production and perception rather than being confined to a single modality. According to Cohn and Schilperoord (2024), who argue for a multimodal language faculty, modality can be understood as a constellation of perceptual, cognitive, and articulatory affordances that enable a semiotic system to function. From this perspective, multimodality is a fundamental property of language itself, as natural languages recruit and integrate signals across multiple modalities. Spoken languages, for example, rely on the vocal modality, yet speech is almost always accompanied by bodily signals such as manual gestures, facial expressions and body movements. Sign languages, by contrast, are realized in the bodily modality through the coordinated use of multiple articulators (hands, face, head, and torso), while also incorporating cross-modal mapping constructions such as mouthing or fingerspelling originating from spoken or written language — see Bauer and Paulus (to appear) on why sign languages are also multimodal.

A second perspective understands multimodality as a property of data. Here, the term refers to the combination of multiple data streams or data gathering techniques used to document language and communication. In this framework, each method — such as audio, video, motion capture, eye-tracking, or EEG — constitutes a distinct modality, and datasets that integrate several of these sources are considered multimodal (Lahat et al. 2015). A key motivation is complementarity: each modality provides unique information that enriches analysis and allows for the study of interdependencies, for example, when eye-tracking data are combined with EEG to investigate fixation-related brain activity (Dimigen & Ehinger

2021). Similarly, audiovisual recordings may be augmented with motion capture to study the coordination of sign and gesture (Puupponen et al. 2015).

Historically, the first perspective emerged from research on sign languages and gesture (Kendon 2004; Sandler & Lillo-Martin 2006; Müller et al. 2013), while the second perspective has become increasingly relevant for the design of multimodal data collection and the technical documentation of communicative behaviour (Pouw et al. 2020; Gregori et al. 2023; Bauer, Trettenbrein et al. 2025). In corpus building, the two perspectives converge: once we acknowledge that language and communication are inherently multimodal, we are compelled to collect richer, multi-layered datasets that can capture their full complexity (which is entirely in line with earlier calls to create multimodal corpora, e.g., Allwood 2008; Abuczki & Baiat Ghazaleh 2013). In this contribution, we argue that uniting these perspectives provides a solid foundation for constructing multimodal corpora and advancing multimodal research in linguistics and beyond.

2.2 Investigating multimodal feedback across languages

In order to truly understand the multimodal nature of language, we suggest that there are two key properties that need to be taken into account. First, we take for granted that in order to truly understand multimodal language use, it is paramount to record and investigate it in the prime habitat of language, namely co-present social interaction, as argued by Seyfeddinipur et al. (this volume). Moreover, a comprehensive understanding of multimodality ultimately requires a cross-linguistic perspective that examines communicative practices across languages to reveal the full range of human linguistic variability.

Throughout this contribution, our reasoning is informed by experiences from our current cross-linguistic research on multimodal recipient feedback in conversation; we occasionally use examples from this research to illustrate a point. Investigating multimodal feedback is crucial for understanding human communication, as feedback significantly contributes to managing attention, building common ground, and signalling problems such as misalignment or breakdowns in understanding (Schegloff 1982; Gardner 2001; Dingemanse & Enfield 2015). While earlier research often concentrated on vocal feedback (Yngve 1970; Jefferson 1984), with video recordings becoming more widely available scholars began extending the scope of analysis to non-vocal signals. Early contributions include Dittmann and Llewellyn (1968) who examined the relation between vocalizations and head nods, and Brunner (1979) who studied smiles as listener feedback. Subsequent research increasingly incorporated visual feedback such as head movements, laughter, facial expressions, and manual gestures (Bavelas et al. 2000; Allwood & Cerrato 2003; Paggio & Navarretta 2013; Hömke et al. 2025). Moreover, recent

research demonstrates that visual cues are pervasive in both sign and spoken languages alike (Lutzenberger et al. 2024).

In our current research, we investigate cross-linguistic aspects of multimodal recipient feedback, aiming to shed light on the similarities and differences between languages. In this contribution, we use examples from two recent or ongoing studies: (1) an investigation of multimodal feedback in dyadic conversation, comparing two spoken and two sign languages matched for cultural background (Bauer, Gipper et al. 2026), and (2) an inquiry into typological aspects of multimodal feedback in a sample of five spoken languages, three of which are non-Indo-European languages (Gipper et al. 2025).

3. Types of language corpora

Language corpora can be classified along many dimensions, including their size, structure, content, and purpose. Given this diversity, our aim here is not to offer a comprehensive typology or to propose criteria for classification. Instead, for the purposes of our argument, we group corpora into three broad types — language documentation corpora, sign language corpora, and majority language corpora. We are aware that many other classifications would be possible and that considerable diversity exists within each category. Nevertheless, this tripartite grouping allows us to highlight a central point: although majority language corpora have been highly influential in linguistic research, they remain underdeveloped in terms of multimodal coverage compared to corpora emerging from language documentation and sign language research. In what follows, we briefly describe each corpus type and outline some typical features, without a claim to exhaustivity. Table 1 summarizes the major points of comparison, ultimately underscoring the limitations of majority language corpora in terms of multimodal richness (see Figure 1).

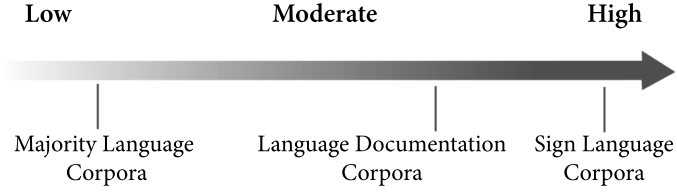


Figure 1. Multimodal richness across language corpus types

Table 1. Key features of the three types of language corpora

Feature	Language documentation corpora	Sign language corpora	Majority language corpora
Primary goal	Create a comprehensive record of communicative practices	Document sign language use	Large-scale analysis
Language focus	Endangered and smaller-scale languages	Sign languages (national, urban, rural, or alternate)	Majority national languages (Chinese, English, German, etc.)
Data modality	Multimodal (audio, often video)	Always multimodal (video)	Mostly written text or audio; limited multimodal content
Data type	Diverse communicative events with different degrees of spontaneity (narratives, conversations, etc.)	Semi-spontaneous staged interactive dyadic signing	If multimodal, often structured formats (interviews, scripted speech)
Size	Smaller to medium-sized	Medium to large	Often large to very large
Recording methodology	Fieldwork; community engagement; variable tech sophistication	Multi-camera setups; advanced video tech; pose tracking	Often TV/studio or structured recordings
Interaction focus	High to moderate, many communicative event types	High, emphasis on conversational data	Lower, often lacks naturalistic interaction
Community orientation	Strong	Strong	Low to moderate

3.1 Language documentation corpora

Language documentation as a field of linguistic inquiry emerged in response to growing concerns among linguists about the accelerating rate of language endangerment and the potential loss of much of their object of investigation, as well as the implications of such loss for the affected communities (Hale 1998). This awareness was notably sparked by Johannes Bechert's presentation at the Fourteenth International Congress of Linguists in East Berlin in 1987 and was accompanied by a broader recognition of the necessity to extend the empirical basis for language research (Himmelmann 2008: 339–340). This resulted in the creation of a wealth of language documentation corpora of endangered and mostly smaller-

scale languages. The data basis for language research has thus indeed grown considerably as a consequence of the *Documentarist Turn*.

Nikolaus Himmelmann was one of the pioneers in defining and shaping the field of language documentation. In his seminal 1998 *Linguistics* paper "Documentary and descriptive linguistics", he advocates for considering language documentation and description as two separate disciplines. Moreover, he formulates his influential definition of the objective of language documentation as creating "a comprehensive record of the linguistic practices characteristic of a given speech community" (Himmelmann 1998: 166). Thus, language documentation as envisioned by Nikolaus Himmelmann (1998) puts the collection of "primary language data" centre stage. According to a data taxonomy proposed by Nikolaus Himmelmann in 2012, these include, next to the recordings constituting the "raw data", transcripts and translations which enable an analysis of the data (Himmelmann 2012: 193–194). In order to ensure capturing the richest information possible, language documentation corpora ideally include video recordings (Himmelmann 1998: 168).

A further crucial component of language documentation corpora is that they should constitute a "lasting, multipurpose record of a language" (Himmelmann 2006: 1). To make a language documentation lasting, archiving the collected data has been a central aspect of language documentation from the start (see Rau, this volume). To give just two examples of the coverage of language documentation archives: According to their websites, PARADISEC (*Pacific and Regional Archive for Digital Sources in Endangered Cultures*, The University of Sydney, Australia) contains materials from over 1,370 languages as of October 2025. The *DobeS Archive* for endangered languages, a subsection of *The Language Archive* (Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands), includes data from major documentation projects from over 100 languages.

The idea that a language documentation corpus should serve many different purposes entails that a broad range of different types of data should be collected, ranging from mythological narratives over procedural texts, grammar and lexicon elicitation sessions, community festivities, to everyday conversations (Himmelmann 1998: 176–183). Ideally, these materials will be useful to a variety of different potential users, both in the language community and in the scientific community. This includes not only what Nikolaus Himmelmann calls "observed communicative events", i.e., events where communicative practices are observed or recorded where the observation is the only intervening factor. Rather, events that are lower on the scale of spontaneity also form an integral part of language documentation, including "staged communicative events" that are performed with the sole purpose of being recorded, such as prompted tellings of mythological narratives and retellings of picture book stories. Crucially for our argument, everyday conver-

sations where participants are invited to meet somewhere for the purpose of the recording belong to the category of staged communicative events. A further event type is elicitation, where language consultants are asked to provide information about their language and linguistic practices (Himmelmann 1998: 185–186). Elicitation sessions are essential to gain insights into parts of the speakers' and signers' "metalinguistic knowledge" that may not surface in observed or staged communicative events, including ethnobiological taxonomies or commentaries about linguistic practices (Himmelmann 1998: 169).

Typically, resources in language documentation projects are limited, which means that not all recorded data can be processed to the same extent. This often results in a situation where language documentation corpora include data at different processing stages forming "concentric circles" (Seyfeddinipur et al., this volume). In this scenario, only a relatively small number of recordings have comprehensive annotations that go beyond the transcription and translation, mostly part-of-speech tagging and morpheme glossing (von Prince & Nordhoff 2020). This could potentially also include annotations added by the researcher for a specific research question, such as multimodal behaviour (see Section 4.1). Ideally, transcriptions and translations should be available for as many recordings as possible, as without them, their utility for research is very limited. However, language documentation corpora often also contain a subset of unprocessed recordings — Himmelmann's (2012) "raw data". The logic here is that the transcribed/translated as well as the more comprehensively annotated data will serve as a foundation for annotating the remaining raw data in the future, either manually or automatically with trained AI (as discussed in Section 4.1 for annotations of meaningful visual signals).

3.2 Sign language corpora

Sign languages are realized in the visual-bodily modality, which makes video-recorded data indispensable and makes them a defining feature of all sign language corpora. Corpus-based approaches in sign language research are relatively recent, having emerged in the early 2000s (Johnston & Schembri 2013; Hochgesang et al. 2023). This delay is largely due to the late recognition of sign languages as natural languages, the need for accessible video technology, and the lack of standardized writing or transcription systems. A key turning point was the development of time-aligned video annotation tools such as ELAN (Wittenburg et al. 2006) and iLex (Hanke 2014), which enabled systematic corpus creation.[1] Today, according to the *Sign Language Dataset Compendium* (Kopf et al. 2024), at least 43 sign language

1. This point is, of course, also relevant for language documentation corpora.

corpora exist worldwide, all of which contain (semi-)spontaneous video recordings with at least partial transcription or translation.

Early large-scale projects such as the *Auslan Corpus* (Australian Sign Language), the *BSL Corpus* (British Sign Language), and the *DGS Corpus* (German Sign Language) recruited hundreds of participants and set standards for sign language corpus design (Johnston 2008; Schembri et al. 2011; Fenlon et al. 2015; Hanke et al. 2020). Increasingly, projects emphasize the importance of involving Deaf participants and signers at all stages of data collection and annotation, thereby strengthening community orientation and ensuring the cultural relevance of corpora.

To enhance data quality and analytical utility, sign language corpus projects have steadily improved their recording technologies. In the *BSL Corpus*, for example, participants were filmed using two high-definition cameras per individual and a third capturing the interaction from the side. The *DGS Corpus* went further by employing five cameras, two of which were mounted above participants to provide a bird's-eye view of the signing space (Hanke et al. 2020; Bauer, Trettenbrein et al. 2025). And although a multi-camera set-up is clearly easier in a lab, it has also found practice in the field (Green 2014).

In this section, we group all sign languages together, fully aware that they constitute a highly diverse set of languages and communicative practices. The corpora discussed above are skewed toward urban national sign languages in countries of the Global North, but sign language typology has also been enriched by research on rural sign languages that emerge in village communities, often in the Global South, where high incidences of hereditary deafness lead to shared signing practices among Deaf and hearing members alike (Nyst 2007; Vos & Zeshan 2012; Satyawati et al. 2025). Another category is constituted by alternate sign languages, developed by hearing populations for specific purposes — such as communication in noisy work environments, religious vows of silence, lingua franca functions, or culturally grounded speech taboos (Kendon 1988; Pfau 2012). Some of these have also been adopted as primary languages by Deaf community members, as documented for Plains Indian Sign Language (Davis 2010) and Yolngu Sign Language in Australia (Bauer 2014).

Sign language corpora often contain semi-spontaneous, interactive data — most commonly dyadic conversations. This is particularly true for corpora of rural sign languages developed with a documentation focus. In contrast, some corpora of national sign languages, such as the *Russian Sign Language (RSL) Corpus* (Burkova 2015), prioritize elicited narrations and therefore contain relatively little interactive material. However, as projects expand, interactive data — primarily consisting of (semi-)spontaneous conversations — increasingly become central. The content is often curated to support linguistic analysis while also doc-

umenting aspects of Deaf culture. *DGS Corpus* participants were, for example, encouraged to share general life experiences, Deaf-specific stories, humour, and personal narratives (Schulder & Hanke 2022). Topics are selected to resonate with the target Deaf communities, ensuring that the resulting resource is not only informative and valuable for research but also engaging, culturally relevant, and supportive of community identity. In this respect, sign language corpora have very similar goals to documentation corpora of endangered languages. Their multimodal and interaction-rich character also brings them closer to language documentation corpora than to the majority language corpora discussed in Section 3.3.

3.3 Majority language corpora

Corpora of highly-resourced and well-described spoken majority languages such as English, German, or Spanish have long served as the primary basis for much of corpus linguistic research. The number of available corpora on majority languages is huge and constantly growing. However, the number of multimodal corpora including both audio and video is extremely low when compared to written and/or audio-recorded spoken corpora. When looking for resources that provide multimodal language use in social interaction, this number further shrinks. Moreover, many such corpora were compiled exclusively from structured formats such as interviews or political debates, or even (semi-)scripted formats such as TV shows or movies, e.g. the *UCLA-Red Hen Lab* multi-billion-word repository of video recordings (Pagán Cánovas et al. 2020).

To illustrate the scarcity of video recordings for majority languages, the largest English-language corpora — the *British National Corpus* (BNC Consortium 2007) and the *Corpus of Contemporary American English* (Davies 2008) — contain spoken texts that account for only 10% and 12.5% of all words, respectively. Crucially, neither corpus includes video recordings. Another example is the *Multimodal Russian Corpus*, part of the *Russian National Corpus* (Savčuk et al. 2024). Although it is described as one of the largest open-access multimodal corpora (Dobrushina & Sokur 2022), the multimodal corpus consists of fragments from films (mostly dating from the 1930s to the 1980s), television and radio talk shows, theatrical performances, and written-to-be-spoken speech (Savčuk & Makhova 2021: 45). While these materials are valuable, and significant research on individual Russian gestures has emerged from this corpus (see Grišina 2017), the fragmented nature of the data and the lack of naturalistic data components makes comprehensive corpus-based multimodal interactional analyses difficult, if not impossible.

But, of course, there are notable exceptions to the written-data bias of majority language corpora. A case in point is the *Datenbank Gesprochenes Deutsch* [Database Spoken German] (DGD). This collection of corpora maintained by the *Insti-*

tut für Deutsche Sprache [Institute for the German Language] (IDS) in Mannheim contains a broad range of resources on spoken German varieties. One of its subcorpora is the *Forschungs- und Lehrkorpus* [Corpus for research and teaching] (FOLK) (Reineke et al. 2023). It contains audio- and video-recorded data of conversational German in a broad range of settings, thus coming close to the ideal of language documentation. Further corpora featuring (relatively) free conversation include the *Nijmegen Corpora of Casual Speech* (Ernestus 2000–2024) which contain conversational data for Dutch, Czech, French, Spanish, as well as Spanish native speakers producing English speech. The *Freiburg Multimodal Interaction Corpus* (FreMIC) is a rich multimodal corpus of free conversation in English containing eye-tracking and pupillometric data, gesture annotation, as well as electrodermal activity data (Rühlemann & Ptak 2023). The *Parallel European Corpus of Informal Interaction* (PECII) includes corpora of naturalistic multimodal language use in a range of European languages, English, Finnish, French, German, Italian, and Polish, documenting three types of activities that ensure comparability of data across corpora: family breakfasts, board game sessions, and car rides (Küttner et al. 2024).

In sum, majority language corpora lag behind in terms of the availability of corpora representing everyday multimodal language use when compared to language documentation and sign language corpora. While corpora of written and scripted data undeniably represent an extremely valuable resource for language scientists, it is increasingly questioned to what extent theories and models of language should rely exclusively on these kinds of samples. This is particularly relevant given that written language lacks crucial aspects of co-present interaction — such as behavioural alignment (Rasenberg et al. 2020), and phenomena related to online language processing (Rühlemann & Barthel 2024).

4. Multimodal corpus building

This contribution argues for a *Multimodal Turn* in corpus building, irrespective of the language being documented. Following Himmelmann's understanding of the goals of language documentation (see Section 3.1), we contend that the standard set for language documentation corpora should be extended to all language corpora: they ought to serve as records that are both maximally accessible and relevant for a wide range of potential users, including members of the language community and their descendants, historians, anthropologists, educators, language planners, and, of course, linguists with diverse theoretical orientations and research interests. As an example of these diverse interests, in this contribution we take the perspective of cross-linguistic investigations of multimodal language use in conversation,

drawing from experiences we made during our research on multimodal recipient feedback mentioned in Section 2.2. In our discussion, which is not meant to be exhaustive, we furthermore build on three key ideas by Nikolaus Himmelmann (1998) that are foundational to language documentation: (1) that it should serve a multitude of potential purposes; (2) that it should constitute a lasting record of a language; and (3) that it focuses on observable events of language use.

4.1 "A multipurpose record": Multimodal data for cross-linguistic research

We argue that the inclusion of video recordings is indispensable for every corpus project, regardless of language or context, as also emphasized by Seyfeddinipur et al. (this volume). Omitting such recordings risks marginalizing the very aspects of language that depend on its situated, embodied, and interactional nature. Only by committing to this multimodal standard in corpus building can we ensure that spoken and signed language is taken seriously as an object of linguistic inquiry.

Interest in multimodal corpora is not new, nor is the call for their systematic development and wider availability (Allwood 2008; Adolphs & Carter 2013; Knight & Adolphs 2020). Indeed, a growing number of linguistic studies have demonstrated the benefits of multimodal data, showing how meaning is co-constructed across modalities — for example, through the interplay of speech and gesture in discourse (Zima 2017; Chen & Adolphs 2023), prosody and manual gesture (Rohrer et al. 2025), or signing and mouthing in sign languages (Bauer & Paulus, to appear). Building on these insights, numerous contributions have also recently addressed the practical aspects of collecting and preparing multimodal data for research (including anonymization, synchronization, storage and archiving of the data), and technical equipment of the highest standard is readily available to many projects today (Pouw et al. 2020; Gregori et al. 2023; Bauer, Trettenbrein et al. 2025).

In our recent research, we were interested in investigating multimodal aspects of recipient feedback in conversation across languages, specifically comparing sign and spoken languages. In order to ensure matching of areal cultural background, we chose two language pairs: German Sign Language (DGS) and spoken German, and Russian Sign Language (RSL) and spoken Russian. Crucially, these data had to be comparable on further dimensions, as conversational dynamics differ with the number of participants and the ongoing activity. Therefore, we chose to work with dyadic conversations where the participants are seated and do not perform any other activity (e.g., eating, cooking) for most of the recording. For spoken German, we used data from a small corpus compiled by Bettina Hoffmann and Nikolaus Himmelmann containing dyadic conversations among fellow students (Hoffmann & Himmelmann 2009). For DGS, we selected three recordings from version 3.0 of the public corpus *MY DGS — Annotated* (Hanke et al.

2020; Konrad et al. 2020). For RSL, we used published as well as unpublished and self-recorded data. One recording was taken from the *RSL Online Corpus*, a corpus compiled at Novosibirsk University (Burkova 2015). The amount of conversational data in this corpus is, however, limited. Therefore, we furthermore used data from a corpus featuring Deaf native signers of RSL recorded by the first author and Roman Poryadin (Bauer & Poryadin 2023). Due to the lack of published corpora containing video-recordings of everyday conversations in Russian (see Section 3.3), our data for spoken Russian also come from a multimodal corpus compiled by the first author (Bauer 2023). This corpus consists of dyadic conversations among native speakers of Russian who immigrated to Germany within the past five years.

Using the data sourced from these corpora, we sampled roughly 1,900 occurrences of feedback. We show that for all languages, non-manual visual signals are involved in over 85% of the cases, with the head emerging as the most pervasively used articulator (Bauer, Gipper et al. 2026). By contrast, vocal-only feedback accounted for just 15% of cases in spoken German and a mere 4% in spoken Russian. Comparable findings were reported for spoken English and British Sign Language by Lutzenberger et al. (2024). This case study illustrates the importance of including visual signals in investigations of communicative behaviour.

But is this "visual bias" in feedback universal? Or is it rather a property of European languages and cultures? In order to investigate this question, we are currently conducting a comparative study of multimodal feedback in a small sample of genealogically diverse languages. Among other sources, we draw on data from documentation corpora, among them *The Yurakaré Archive* (van Gijn et al. 2011). This corpus stems from a major language documentation carried out by Rik van Gijn, Vincent Hirtzel, the Yurakaré speaker and language expert Jeremías Ballivián Torrico, and the second author. It includes recordings of dyadic everyday conversations and thus offers data that are fully comparable to the data used in Bauer, Gipper et al. (2026). Our preliminary findings point to a clear cross-linguistic contrast in how multimodal feedback is realized: While speakers of the Indo-European languages in our sample rely heavily on non-manual signals, particularly head nods, speakers of the non-Indo-European languages favour verbal means in shaping their feedback events.

The two examples from our research illustrate how existing language documentation and sign language corpora can be fruitfully employed for comparative investigations of multimodal interaction. But the task of selecting a set of primary data is non-trivial. In particular the first example shows that, as in the case of Russian, multimodal video-recorded data may be less readily available for majority languages. But the examples also draw our attention to the comparability of data sourced from corpora for cross-linguistic multimodal research. All the

recordings we employed feature "staged communicative events" where participants were invited to take part in language recording activities for the purpose of the recording. In our experience, such data are extremely valuable for comparative research, as they allow for some control of the setting (in our case, seated dyadic conversation without additional activities) while still representing naturalistic language use.

4.2 "A lasting record": Archiving as an integral part of language research

One of the central insights of Nikolaus Himmelmann's seminal work is that corpora should provide a lasting record of language use, enabling future generations to explore a broad range of research questions. Consequently, archiving raw and primary data is an integral part of language documentation (Himmelmann 1998: 171, 2006: 1–2). In contrast, a crucial problem with multimodal corpora of majority languages is that access is often restricted. This is, for instance the case when data sharing is achieved via informal channels rather than by depositing in a formal repository (Joyce et al. 2022: 653). We argue that archiving practice should also become standard for multimodal data of majority languages.

For endangered languages, the point that language documentation corpora need to be lasting is particularly important, as further data collection may no longer be possible in the future. One may argue that archiving is not strictly necessary in the case of spoken majority languages, as new data can be collected at any time in the future. However, archiving does not only serve the function of preserving language data; it is also a vehicle of making research data accessible to others, in line with recent developments in research reproducibility and Open Science. There is also another argument for archiving language data from majority languages: While it will indeed very likely be possible to collect new data for most majority languages in the future, these languages will probably have changed in the meantime. So with time passing, the archived data will become part of a historical record of a language, making it eventually possible to conduct diachronic investigations of (multimodal) language use (see also Gipper et al., this volume).

But not only raw and primary data may be useful to researchers asking novel questions. From the perspective of investigating multimodal language use across languages, the primary data need further processing in terms of annotations of multimodal activities such as manual gestures, head nods, or facial expressions. One approach is to create manual annotations of bodily behaviour, as in our investigations of multimodal feedback (Bauer, Gipper et al. 2026; Gipper et al. 2025). Another common way of creating annotations of bodily behaviour is through pose estimation, which detects and tracks the position and orientation

of a person's body. Each recording in the public *DGS Corpus* is accompanied by pose estimation data (Schulder & Hanke 2020) extracted post hoc using OpenPose (Cao et al. 2021), a real-time multi-person keypoint detection framework that estimates 2D skeletal positions including the body, hands, and face from RGB video input. This is followed by post-processing to correct common errors. The availability of such systematically corrected pose estimation data enables fine-grained analyses of the kinematic features of multimodal communication, as demonstrated by Bauer et al. (2024).

Looking ahead, it will become increasingly feasible to combine manual and automatic annotation to train AI systems capable of distinguishing meaningful movements (identified through manual annotation) from non-meaningful ones (detected by pose estimation alongside meaningful gestures) (Pouw 2024). Within the internal architecture of language documentation and other multimodal corpora, even a limited subset of manually annotated meaningful movements may prove highly valuable, as these can serve as training data for automated annotation of the corpus as a whole. Consequently, when researchers generate such annotations — whether manually or automatically — it is crucial that they be made accessible alongside the primary data. Here, the value of archiving in multimodal corpus-building becomes evident one more time.

4.3 Beyond the "observable": Incorporating physiological data

In language documentation, the focus of attention clearly lies on observable linguistic behaviour (Himmelmann 1998:166, 2006:7, 2008:346), a feature that is shared by many language corpora in general. In this sense, "observable" refers to events occurring outside the control of the researcher. But it has a second sense, and in this section, we explore potential avenues for integrating features that are not directly observable, in the form of physiological measures of quantifiable biological responses. These techniques, we argue, considerably augment the linguistic information contained in a corpus and open up possibilities for research questions that could not previously be addressed. Recent advancements in mobile and wireless technology enable linguists to integrate physiological measures (for reviews, see Gregori et al. 2023; Hardacre 2020) such as eye-tracking, electroencephalography (EEG), heart rate variability, electrodermal activity, skin temperature, and breath tracking in naturalistic settings rather than confined laboratory environments (e.g., Rochet-Capellan & Fuchs 2014; Dikker et al. 2021; P. Chen et al. 2022). For a discussion of challenges and opportunities in implementing physiological measures in field settings, see Sauppe et al. (2023) as well as Næss and Sauppe (2025).

Our aim is not to suggest that such measures should become a universal standard in multimodal corpus building, but rather to highlight the possibilities they offer and the kinds of research questions they may help to address. Enhanced data including physiological measures fit very well into the internal structure of language documentation corpora, falling within the subset of data that show high degrees of processing including special-purpose information. In what follows, we discuss two possible measures in more detail: eye-tracking and EEG.

Tracking the movements of someone else's eyes is a very natural behaviour that humans constantly perform in social interaction (Stephenson et al. 2021). But eye movements also correlate with language comprehension, as realized in the 1980s (Just & Carpenter 1980). Linguists and psycholinguists have used this correlation between "looking at something" and "processing that something" in reading studies (for a historical overview, see Płużyczka 2018) and other experimental paradigms (Conklin et al. 2018).

Including eye-tracking data in multimodal language corpora will allow to address a range of research questions not possible to tackle without these data. These research questions may pertain to the realm of online processing, but also concern the interaction between a person and their communication partner or environment. Since gaze behaviour varies across cultures (Rossano et al. 2009), analysing participants' eye gaze and its influence on ongoing interaction illustrates how physiological data can enrich multimodal language documentation by providing precise information about participants' gaze patterns during recorded activities. Further applications include, for example, investigations of the functions of intonation and prosody (Wiener & Ito 2014), and studies of the structure and processing of language (Sarvasy et al. 2023). In our current research, we are extending our investigation to the role of gaze in multimodal feedback across languages. As video recordings alone do not allow for a precise analysis of gaze behaviour (Zima et al. 2025), we are currently recording dyadic conversational data with German speakers including eye-tracking with mobile eye-tracking headbands (see Figure 2).

While the first mobile head-mounted eye-trackers were still a rather bulky frame with a mirror and two cameras attached to a helmet (as in Hosemann 2013), the last decades saw the development of smaller and more practical eye-tracking devices that can be worn as glasses or headbands, as shown in Figure 2. This enables researchers to include eye-tracking in lab recordings of casual language use as in the *Freiburg Multimodal Interaction Corpus* of conversations in English (Rühlemann & Ptak 2023), but also in real-life settings in the field such as joint walks (Auer & Laner 2025). Given their suitability for field settings, mobile eye-tracking devices can be integrated in language documentation and other corpus-building projects. This will allow, for instance, cross-linguistic investigations of

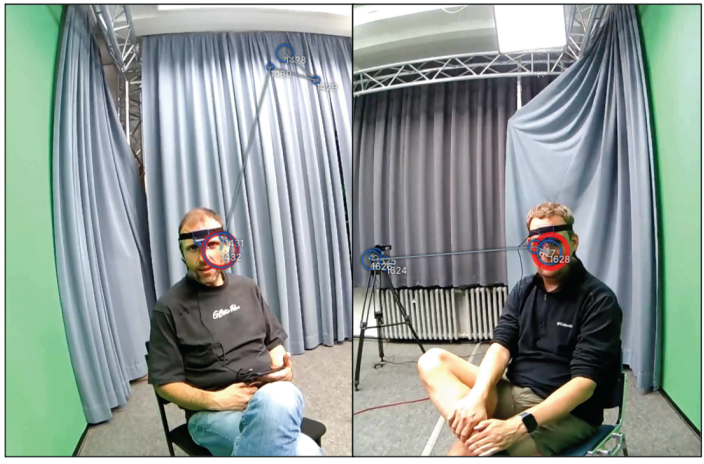


Figure 2. Two German speakers wearing eye-tracking headbands during a recording; cursor positions show their gaze movements

language processing and conversational dynamics, and can also be useful in field elicitation, for instance in ethnobiological studies. Another data source that can be gathered with eye-trackers is the measurement of pupillometric data, i.e., changes in pupil sizes, which can provide insights into cognitive load processes during interaction (Rühlemann & Barthel 2024).

A further physiological measure that seems promising for inclusion in multimodal corpora is EEG, a method for continuously capturing electrical activity in the brain which displays fluctuations over time. A specific application of EEG is the observation of event-related potentials (ERPs), i.e., distinct segments of EEG activity corresponding to specific events, such as expectation mismatches caused by words or sounds. ERPs enable researchers to pinpoint precise neural responses to linguistic and cognitive tasks by time-locking stimuli and analysing brain activity at specific moments (e.g., Rees et al. 2019).

Traditionally, EEG requires the participant to put on a cap to which electrodes are attached. Fitting and attaching the electrodes to the head can take about 1–2 hours. Afterwards, the participant is generally sent to an isolated chamber to be presented with the experimental stimuli so that no electrical disturbance or other distractions affect them in order to obtain a clear signal of brain activity. Thanks to modern portable and wireless EEG devices that enable data collection in naturalistic settings, language documentation and multimodal corpus building in general can nowadays expand their reach to the cognitive and emotional states related to language use. The available devices vary in design, measurement capabilities and affordability. Two examples used in research settings are Emotiv EPOC (Badcock et al. 2013; Williams et al. 2020) and InteraXon Muse, a head-

band with built-in sensors along its internal surface (Krigolson et al. 2017; Negi & Mitra 2018; Hayes & Magne 2024).

Recent studies showed that EEG data can also be used to investigate naturalistic human interaction (Mazzini et al. 2023). Instead of comparing event-related potentials in different conditions, here EEG data of two persons are recorded during a conversation. These data can be analysed with intra-brain spectral and time-frequency analyses, or with connectivity and coherence analyses. An overview of so-called EEG-based hyperscanning of social interactions is given in Liu et al. (2018). Thus, there is a promising field of research that can provide new physiological evidence for phenomena such as convergence among interlocutors (Murata et al. 2020), or stress, emotions, and social dynamics shaping conversations (Kroczek et al. 2024).

That said, human brains are constantly active, and EEG waves result from all possible brain activities. Therefore, it is extremely difficult to connect a certain event within a conversation to a specific EEG signal. Nevertheless, analysing EEG data during natural human conversation and in language documentation can give us information about higher or lower (cognitive) engagements, and a higher or lower cortical tracking of speech or bodily signals (Mazzini et al. 2023). Thus, we propose that there are many promising avenues for including EEG recordings in language corpora. Further advances in the understanding of the brain may allow a more sophisticated interpretation of the brain activity recordings in the future that go beyond questions we can ask today.

Apart from eye-tracking and EEG, a non-exhaustive list of physiological measures that correlate with cognitive and emotional states and can be collected using portable devices includes heart rate variability, electrodermal activity, skin temperature, and breathing rhythm. Many fitness watches are now equipped to track these. An example of a watch used for research applications is the Empatica E4 bracelet (Borrego et al. 2019; Kozanitis 2023). Such physiological measures have also been successfully applied to dyadic interactions to assess interpersonal engagement using electrocardiography, skin conductance, respiration and skin temperature (Chatterjee et al. 2021; Offrede et al. 2024), gesture expressivity and emotional resonance in conversational storytelling including electrodermal activity (Rühlemann & Trujillo 2024), emotional arousal and its non-verbal correlates using galvanic skin response paired with a gesture analysis (Lindberg et al. 2021), and synchronization among interlocutors — see Marzoratti & Evans (2022) and Helm et al. (2018) for reviews on measuring physiological synchrony in conversation, including the use of electrodermal skin activation, cardiac function, respiration, and neural oscillatory activity.

5. Concluding remarks

For decades, underresourced spoken and sign languages lacked the corpora needed for systematic investigation, and consequently were underrepresented in linguistic research. This imbalance is now shifting. Some 65 years after the recognition of sign languages as natural languages (1960), 40 years after Bechert's influential 1987 talk, and 30 years after Himmelmann's seminal 1998 paper, language documentation corpora now provide a rich foundation for investigating the multimodal nature of language. The *Documentarist Turn* thus constitutes a central pillar of the emerging *Multimodal Turn* in the language sciences. Similarly, many sign languages are documented in richly conversational corpora with HD video, providing unique access to the interactional dimensions of language. In comparison, majority languages still face a shortage of resources that capture spontaneous, multimodal interaction. Taking inspiration from these corpus-building traditions, we suggested in this contribution that future corpus-building efforts should: (1) make video-recording of a broad variety of communicative event types the standard for data collection in order to enable multimodal research; (2) archive the collected data to preserve them for the future and make them accessible to others; (3) potentially include physiological measures that are not directly observable in order to capture the full richness of multimodal language use in (inter)action.

By foregrounding the dimension of multimodality and highlighting the broader trend in modern linguistics toward interactive aspects of language use, we argue that corpus building — particularly for majority languages — must learn from the paradigms of language documentation and sign language research. Doing so is not only essential to filling a methodological gap but also to transforming the future of linguistic inquiry itself. With this contribution, we seek to reinforce the perspective that corpora should be designed to enable multimodal research, both on particular languages as well as across languages. If we are to truly understand language as it lives in human interaction, corpus building must resolutely turn toward the multimodal.

Funding

This research was funded by the University of Cologne *Excellent Research Program*, Funding Line *Cluster Development Program*, Project *Language Challenges*.

Acknowledgments

We are deeply grateful to Birgit Hellwig and two anonymous reviewers for their inspiring comments on an earlier version of this manuscript; their feedback has significantly strengthened and sharpened this contribution.

Above all, we feel very fortunate to have Nikolaus Himmelmann as a colleague and mentor. His generosity, vision, and constant encouragement have inspired us throughout the last project and far beyond. For the first two authors in particular, his mentorship has been a defining influence, shaping not only their development as linguists but also their confidence to pursue ambitious research.

Author statement

The first two authors contributed equally to the conception, design, and preparation of the manuscript and share first authorship. The remaining authors contributed in a supporting capacity and are listed in alphabetical order.

References

doi Abner, Natasha, Kensy Cooperrider, and Susan Goldin-Meadow. 2015. "Gesture for Linguists: A Handy Primer." *Language and Linguistics Compass* 9 (11): 437–451.

Abuczki, Ágnes, and Esfandiari Baiat Ghazaleh. 2013. "An Overview of Multimodal Corpora, Annotation Tools and Schemes." *Argumentum* 9: 86–98.

doi Adolphs, Svenja, and Ronald Carter. 2013. *Spoken Corpus Linguistics: From Monomodal to Multimodal*. New York and London: Routledge.

doi Allwood, Jens. 2008. "Multimodal Corpora." In *Corpus Linguistics: An International Handbook*, Vol. 1, ed. by Anke Lüdeling, and Merja Kytö, 207–225. Berlin: De Gruyter Mouton.

Allwood, Jens, and Loredana Cerrato. 2003. "A Study of Gestural Feedback Expressions." In *First Nordic Symposium on Multimodal Communication*, ed. by Patrizia Paggio, Kristiina Jokinen, and Arne Jönsson, 7–22. Copenhagen: Gothenburg University Publications.

doi Ameka, Felix K., and Marina Terkourafi. 2019. "What If…? Imagining Non-Western Perspectives on Pragmatic Theory and Practice." *Journal of Pragmatics* 145: 72–82.

doi Auer, Peter, and Barbara Laner. 2025. "Laughter and Gaze among Talkers on a Walk." In *Mobile Eye Tracking: New Avenues for the Study of Gaze in Social Interaction*, ed. by Elisabeth Zima, and Anja Stukenbrock, 208–242. Amsterdam: John Benjamins.

doi Badcock, Nicholas A., Petroula Mousikou, Yatin Mahajan, Peter de Lissa, Johnson Thie, and Genevieve McArthur. 2013. "Validation of the Emotiv EPOC® EEG Gaming System for Measuring Research Quality Auditory ERPs." *PeerJ* 1: e38.

doi Bauer, Anastasia. 2014. *The Use of Signing Space in a Shared Sign Language of Australia*. Berlin: De Gruyter Mouton.

Bauer, Anastasia. 2023. *Russian Multimodal Conversational Corpus*. Data Center for the Humanities, University of Cologne.

Bauer, Anastasia, Sonja Gipper, Tobias-Alexander Herrmann, and Jana Hosemann. 2026. "Rethinking Linguistic Feedback: A Modality-Agnostic and Holistic Approach to Multimodal Addressee Signals in Spoken and Signed Dyadic Interaction." *Glossa: a journal of general linguistics* 11 (1).

Bauer, Anastasia, Anna Kuder, Marc Schulder, and Job Schepens. 2024. "Phonetic Differences between Affirmative and Feedback Head Nods in German Sign Language (DGS): A Pose Estimation Study." *PLoS ONE* 19 (5): e0304040.

Bauer, Anastasia, and Liona Paulus. To appear. "The Multimodality of Sign Languages." In *The Handbook of Linguistics and Multimodality*, ed. by John Bateman, and Chiao-I. Tseng. Hoboken, NJ: Wiley.

Bauer, Anastasia, and Roman Poryadin. 2023. *Russian Sign Language Conversations*. Data Center for the Humanities, University of Cologne.

Bauer, Anastasia, Patrick C. Trettenbrein, Federica Amici, Aleksandra Ćwiek, Susanne Fuchs, Lisa-Marie Krause, Anna Kuder, Silva Ladewig, Marc Schulder, Petra Schumacher, Door Spruijt, Chiara Zulberti, and Martin Schulte-Rüther. 2025. "Data Collection in Multimodal Language and Communication Research: A Flexible Decision framework." Preprint, OSF.

Bavelas, Janet, Nicole Chovil, Linda Coates, and Lori Roe. 1995. "Gestures Specialized for Dialogue." *Personality and Social Psychology Bulletin* 21 (4): 394–405.

Bavelas, Janet, Linda Coates, and Trudy Johnson. 2000. "Listeners as Co-Narrators." *Journal of Personality and Social Psychology* 79 (6): 941–952.

BNC Consortium. 2007. *The British National Corpus XML Edition*. http://hdl.handle.net/20.500.14106/2554

Borrego, Adrian, Jorge Latorre, Mariano Alcaniz, and Roberto Llorens. 2019. "Reliability of the Empatica E4 Wristband to Measure Electrodermal Activity to Emotional Stimuli." In *Proceedings of the 2019 International Conference on Virtual Rehabilitation (ICVR)*, 1–2. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Brunner, Lawrence J. 1979. "Smiles Can Be Back Channels." *Journal of Personality and Social Psychology* 37 (5): 728–734.

Burkova, Svetlana. 2015. *Russian Sign Language Corpus*. Novosibirsk University. http://rsl.nstu.ru/

Cao, Zhe, Ginés Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A. Sheikh. 2021. "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (1): 172–186.

Chafe, Wallace. 1992. "The Importance of Corpus Linguistics to Understanding the Nature of Language." In *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991*, ed. by Jan Svartvik, 79–104. Berlin: De Gruyter Mouton.

Chatterjee, Iman, Maja Goršič, Joshua D. Clapp, and Domen Novak. 2021. "Automatic Estimation of Interpersonal Engagement during Naturalistic Conversation Using Dyadic Physiological Measurements." *Frontiers in Neuroscience* 15: 757381.

Chen, Phoebe, Ulrich Kirk, and Suzanne Dikker. 2022. "Trait Mindful Awareness Predicts Inter-Brain Coupling but Not Individual Brain Responses during Naturalistic Face-to-Face Interactions." *Frontiers in Psychology* 13: 915345.

Chen, Yaoyao, and Svenja Adolphs. 2023. "Towards a Corpus-Based Description of Speech-Gesture Units of Meaning: The Case of the Circular Gesture." *International Journal of Corpus Linguistics* 28 (2): 172–201.

Cohn, Neil, and Joost Schilperoord. 2024. *A Multimodal Language Faculty: A Cognitive Framework for Human Communication*. London, New York and Dublin: Bloomsbury.

Conklin, Kathy, Ana Pellicer-Sánchez, and Gareth Carrol. 2018. *Eye-Tracking: Applied Linguistics and Second Language Acquisition*. Cambridge: Cambridge University Press.

Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Davis, Jeffrey. 2010. *Hand Talk: Sign Language Among American Indian Nations*. Cambridge: Cambridge University Press.

Dikker, Suzanne, Georgios Michalareas, Matthias Oostrik, Amalia Serafimaki, Hasibe Melda Kahraman, Marijn E. Struiksma, and David Poeppel. 2021. "Crowdsourcing Neuroscience: Inter-Brain Coupling during Face-to-Face Interactions Outside the Laboratory." *NeuroImage* 227: 117436.

Dimigen, Olaf, and Benedikt V. Ehinger. 2021. "Regression-Based Analysis of Combined EEG and Eye-Tracking Data: Theory and Applications." *Journal of Vision* 21 (1): 3.

Dingemanse, Mark, and Nick J. Enfield. 2015. "Other-Initiated Repair across Languages: Towards a Typology of Conversational Structures." *Open Linguistics* 1 (1): 96–118.

Dittmann, Allen T., and Lynn G. Llewellyn. 1968. "Relationship between Vocalizations and Head Nods as Listener Responses." *Journal of Personality and Social Psychology* 9 (1): 79–84.

Dobrushina, Nina, and Elena Sokur. 2022. "Spoken Corpora of Slavic Languages." *Russian Linguistics* 46 (2): 77–93.

Enfield, Nick J. 2009. *The Anatomy of Meaning: Speech, Gesture, and Composite Utterances*. Cambridge: Cambridge University Press.

Ernestus, Mirjam. 2000–2024. *Nijmegen Corpora of Casual Speech*. The Language Archive, MPI Nijmegen. https://hdl.handle.net/1839/eb85d4af-42f3-45f6-84eb-b80c6d11d31c

Fenlon, Jordan, Adam Schembri, Trevor Johnston, and Kearsy Cormier. 2015. "Documentary and Corpus Approaches to Sign Language Research." In *Research Methods in Sign Language Studies*, ed. by Eleni Orfanidou, Bencie Woll, and Gary Morgan, 156–172. Chichester: Wiley Blackwell.

Ferrara, Lindsay, and Gabrielle Hodge. 2018. "Language as Description, Indication, and Depiction." *Frontiers in Psychology* 9: 716.

Gardner, Rod. 2001. *When Listeners Talk: Response Tokens and Listener Stance*. Amsterdam: John Benjamins.

Gijn, Rik van, Vincent Hirtzel, Sonja Gipper, and Jeremías Ballivián Torrico. 2011. *The Yurakaré Archive*. The Language Archive, MPI Nijmegen. https://hdl.handle.net/1839/8df587ed-3d6e-4db8-bfe5-4ecad5cef3a2

Gipper, Sonja, Eugen Hill, and Pascal Coenen. This volume. "Language Change in Real Time: Investigating Conversational Priming in Repetitional Responses." In *The Documentarist Turn: From Observable Linguistic Behaviour to Typological Generalizations*, ed. by Sonja Riesberg, Uta Reinöhl, and Birgit Hellwig. Amsterdam: John Benjamins.

Gipper, Sonja, Anastasia Bauer, Alice Mitchell, and Alena Witzlack-Makarevich. 2025. "Cross-Linguistic Variation in Multimodal Feedback in Conversation." Presentation, 24. Arbeitstagung zur Gesprächsforschung, Mannheim, 19–21 March 2025.

Green, Jennifer. 2014. *Drawn from the Ground: Sound, Sign and Inscription in Central Australian Sand Stories*. Cambridge: Cambridge University Press.

Gregori, Alina, Federica Amici, Ingmar Brilmayer, Aleksandra Ćwiek, Lennart Fritzsche, Susanne Fuchs, Alexander Henlein, Oliver Herbort, Frank Kügler, Jens Lemanski, Katja Liebal, Andy Lücking, Alexander Mehler, Kim Tien Nguyen, Wim Pouw, Pilar Prieto, Patrick Louis Rohrer, Paula G. Sánchez-Ramón, Martin Schulte-Rüther, Petra B. Schumacher, Stefan R. Schweinberger, Volker Struckmeier, Patrick C. Trettenbrein, and Celina I. von Eiff. 2023. "A Roadmap for Technological Innovation in Multimodal Communication Research." In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, ed. by Vincent G. Duffy, 402–438. Cham: Springer.

Grišina, Elena. 2017. *Russkaja žestikuljacija s lingvističeskoj točki zrenija: Korpusnye issledovanija* [*Russian gesticulation from a linguistic point of view: Corpus-based studies*]. Moskva: Izdatel'skij dom JASK (Razumnoe povedenie).

Hagoort, Peter, and Aslı Özyürek. 2025. "Extending the Architecture of Language from a Multimodal Perspective." *Topics in Cognitive Science* Online First: 12728.

Hale, Ken. 1998. "On Endangered Languages and the Importance of Linguistic Diversity." In *Endangered Languages: Language Loss and Community Response*, ed. by Lenore A. Grenoble, and Lindsay J. Whaley, 192–216. Cambridge: Cambridge University Press.

Hamilton, Antonia F. de C., and Judith Holler. 2023. "Face2face: Advancing the Science of Social Interaction." *Philosophical Transactions of the Royal Society B: Biological Sciences* 378 (1875): 20210470.

Hanke, Thomas. 2014. "Annotation of Mouth Activities with iLex." In *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel (LREC 2014)*, ed. by Onno Crasborn, Eleni Efthimiou, Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch, 67–70. Paris: European Language Resources Association.

Hanke, Thomas, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. "Extending the Public DGS Corpus in Size and Depth." In *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives (LREC 2020)*, ed. by Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch, 75–82. Paris: European Language Resources Association.

Hardacre, Bahiyyih. 2020. *Psychophysiological Methods in Language Research: Rethinking Embodiment in Studies of Linguistic Behaviors*. Lanham, MD: Lexington.

Hayes, Hannah Begue, and Cyrille Magne. 2024. "Exploring the Utility of the Muse Headset for Capturing the N400: Dependability and Single-Trial Analysis." *Sensors* 24 (24): 7961.

Helm, Jonathan L., Jonas G. Miller, Sarah Kahle, Natalie R. Troxel, and Paul D. Hastings. 2018. "On Measuring and Modeling Physiological Synchrony in Dyads." *Multivariate Behavioral Research* 53 (4): 521–543.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–196.

Himmelmann, Nikolaus P. 2006. "Language Documentation: What Is It and What Is It Good for? In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2008. "Reproduction and Preservation of Linguistic Knowledge: Linguistics' Response to Language Endangerment." *Annual Review of Anthropology* 37 (1): 337–350.

Himmelmann, Nikolaus P. 2012. "Linguistic Data Types and the Interface between Language Documentation and Description." *Language Documentation and Conservation* 6: 187–207. http://hdl.handle.net/10125/4503

Hochgesang, Julie A., Ryan Lepic, and Emily Shaw. 2023. "W(h)ither the ASL Corpus? Considering Trends in Signed Corpus Development." In *Advances in Sign Language Corpus Linguistics*, ed. by Ella Wehrmeyer, 287–308. Amsterdam: John Benjamins.

Hoffmann, Bettina, and Nikolaus P. Himmelmann. 2009. "Münster Videokorpus Alltagsgespräche [Münster video corpus of everyday conversations]." Unpublished corpus.

Holler, Judith, and Stephen C. Levinson. 2019. "Multimodal Language Processing in Human Communication." *Trends in Cognitive Sciences* 23 (8): 639–652.

Hömke, Paul, Stephen C. Levinson, Alexandra K. Emmendorfer, and Judith Holler. 2025. "Eyebrow Movements as Signals of Communicative Problems in Human Face-to-Face Interaction." *Royal Society Open Science* 12: 12241632.

Hosemann, Jana. 2013. "Eye Gaze and Verb Agreement in German Sign Language: A First Glance." In *Nonmanuals in Sign Language*, ed. by Annika Herrmann, and Markus Steinbach, 73–90. Amsterdam: John Benjamins.

Jefferson, Gail. 1984. "Notes on a Systematic Deployment of the Acknowledgement Tokens 'Yeah'; and 'Mm Hm';" *Paper in Linguistics* 17 (2): 197–216.

Johnston, Trevor. 2008. *Auslan Corpus*. Endangered Languages Archive. http://hdl.handle.net/2196/00-0000-0000-0000-D7CF-8

Johnston, Trevor, and Adam Schembri. 2013. "Corpus Analysis of Sign Languages." In *The Encyclopedia of Applied Linguistics*, ed. by Carol A. Chapell, 479–501. Chichester: Wiley Blackwell.

Joyce, Jack B., Tom Douglass, Bethan Benwell, Catrin S. Rhys, Ruth Parry, Richard Simmons, and Adrian Kerrison. 2022. "Should We Share Qualitative Data? Epistemological and Practical Insights from Conversation Analysis." *International Journal of Social Research Methodology* 26 (6): 645–659.

Just, Marcel A., and Patricia A. Carpenter. 1980. "A Theory of Reading: From Eye Fixations to Comprehension." *Psychological Review* 87 (4): 329–354.

Keevallik, Leelo. 2018. "What Does Embodied Interaction Tell Us about Grammar?" *Research on Language and Social Interaction* 51 (1): 1–21.

Kendon, Adam. 1988. *Sign Languages of Aboriginal Australia: Cultural, Semiotic and Communicative Perspectives*. Cambridge: Cambridge University Press.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kendrick, Kobin H., Judith Holler, and Stephen C. Levinson. 2023. "Turn-Taking in Human Face-to-Face Interaction Is Multimodal: Gaze Direction and Manual Gestures Aid the Coordination of Turn Transitions." *Philosophical Transactions of the Royal Society B: Biological Sciences* 378 (1875): 20210473.

Knight, Dawn, and Svenja Adolphs. 2020. "Multimodal Corpora." In *A Practical Handbook of Corpus Linguistics*, ed. by Magali Paquot, and Stefan Th. Gries, 353–371. Cham: Springer.

Konrad, Reiner, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, and Marc Schulder. 2020. *MY DGS — Annotated. Public Corpus of German Sign Language*. Release 3.0.

Kopf, Maria, Marc Schulder, and Thomas Hanke. 2024. *The Sign Language Dataset Compendium*. Tech. rep. Version 1.3.1. Hamburg University.

Kozanitis, Anastassis. 2023. "Comparing Real Time Student Situational Engagement in Traditional and Active Learning Classroom Using Non-Invasive Electrodermal Measurements." In *Proceedings of the 14th International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC 2023)*, ed. by Nagib C. Callaos, Shigehiro Hashimoto, Natalja Lace, Belkis Sánchez, and Michael Savoie, 133–138. Winter Garden: International Institute of Informatics and Cybernetics.

Krigolson, Olave E., Chad C. Williams, Angela Norton, Cameron D. Hassall, and Francisco L. Colino. 2017. "Choosing MUSE: Validation of a Low-Cost, Portable EEG System for ERP Research." *Frontiers in Neuroscience* 11: 109.

Kroczek, Leon O. H., Angelika Lingnau, Valentin Schwind, Christian Wolff, and Andreas Mühlberger. 2024. "Observers Predict Actions from Facial Emotional Expressions during Real-Time Social Interactions." *Behavioural Brain Research* 471: 115126.

Küttner, Uwe-A., Laurenz Kornfeld, Christina Mack, Lorenza Mondada, Jowita Rogowska, Giovanni Rossi, Marja-Leena Sorjonen, Matylda Weidner, and Jörg Zinken. 2024. "Introducing the 'Parallel European Corpus of Informal Interaction' (PECII): A Novel Resource for Exploring Cross-Situational and Cross-Linguistic Variability in Social Interaction." In *New Perspectives in Interactional Linguistic Research*, ed. by Margret Selting, and Dagmar Barth-Weingarten, 132–160. Amsterdam: John Benjamins.

Lahat, Dana, Tülay Adali, and Christian Jutten. 2015. "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects." *Proceedings of the IEEE* 103 (9): 1449–1477.

Levinson, Stephen C. 2006. "On the Human "Interaction Engine"." In *Roots of Human Sociality: Culture, Cognition and Interaction*, ed. by Nick J. Enfield, and Stephen C. Levinson, 39–69. Oxford: Berg.

Lindberg, Rachael, Kim McDonough, and Pavel Trofimovich. 2021. "Investigating Verbal and Nonverbal Indicators of Physiological Response during Second Language Interaction." *Applied Psycholinguistics* 42 (6): 1403–1425.

Liu, Difei, Shen Liu, Xiaoming Liu, Chong Zhang, Aosika Li, Chenggong Jin, Yijun Chen, Hangwei Wang, and Xiaochu Zhang. 2018. "Interactive Brain Activity: Review and Progress on EEG-based Hyperscanning in Social Interactions." *Frontiers in Psychology* 9: 1862.

Lutzenberger, Hannah, Lierin de Wael, Rehana Omardeen, and Mark Dingemanse. 2024. "Interactional Infrastructure across Modalities: A Comparison of Repair Initiators and Continuers in British Sign Language and British English." *Sign Language Studies* 24 (3): 548–581.

Macuch Silva, Vinicius, Judith Holler, Aslı Özyürek, and Seán G. Roberts. 2020. "Multimodality and the Origin of a Novel Communication System in Face-to-Face Interaction." *Royal Society Open Science* 7 (1): 182056.

Marzoratti, Analia, and Tanya M. Evans. 2022. "Measurement of Interpersonal Physiological Synchrony in Dyads: A Review of Timing Parameters Used in the Literature." *Cognitive, Affective, and Behavioral Neuroscience* 22 (6): 1215–1230.

Mazzini, Sara, Judith Holler, and Linda Drijvers. 2023. "Studying Naturalistic Human Communication Using Dual-EEG and Audio-Visual Recordings." *STAR Protocols* 4 (3): 102370.

Mondada, Lorenza. 2016. "Challenges of Multimodality: Language and the Body in Social Interaction." *Journal of Sociolinguistics* 20 (3): 336–366.

Müller, Cornelia, Silva H. Ladewig, and Jana Bressem. 2013. "Gestures and Speech from a Linguistic Perspective: A New Field and Its History." In *Body — Language — Communication: An International Handbook on Multimodality in Human Interaction*, ed. by Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, and Sedinha Tessendorf, 55–81. Berlin: De Gruyter Mouton.

Murata, Aiko, Hiroshi Nishida, Katsumi Watanabe, and Tatsuya Kameda. 2020. "Convergence of Physiological Responses to Pain during Face-to-Face Interaction." *Scientific Reports* 10 (1): 450.

Næss, Åshild, and Sebastian Sauppe. 2025. "Bringing Psycholinguistics to the Field: Experiences from Solomon Islands." *Language Documentation and Conservation* 19: 99–119. https://hdl.handle.net/10125/74816

Negi, Shivsevak, and Ritayan Mitra. 2018. "EEG Metrics to Determine Cognitive Load and Affective States: A Pilot Study." In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp 2018)*, 182–185. New York: Association for Computing Machinery.

Nyst, Victoria. 2007. *A Descriptive Analysis of Adamorobe Sign Language (Ghana)*. Utrecht: LOT. https://hdl.handle.net/11245/1.266338

Offrede, Tom, Christine Mooshammer, Alessandro D'Ausilio, and Susanne Fuchs. 2024. "Speech Adaptation and Physiological Responses: A Study on f0 and Skin Temperature." In *Proceedings of the International Conference on Speech Prosody 2024*, ed. by Yiya Chen, Aoju Chen, and Amalia Arvaniti, 467–471. International Speech Communication Association.

Özyürek, Aslı. 2021. "Considering the Nature of Multimodal Language from a Crosslinguistic Perspective." *Journal of Cognition* 4 (1): 42.

Pagán Cánovas, Cristóbal, Javier Valenzuela, Daniel Alcaraz Carrión, Inés Olza, and Michael Ramscar. 2020. "Quantifying the Speech-Gesture Relation with Massive Multimodal Datasets: Informativity in Time Expressions." *PLoS ONE* 15 (6): e0233892.

Paggio, Patrizia, and Costanza Navarretta. 2013. "Head Movements, Facial Expressions and Feedback in Conversations: Empirical Evidence from Danish Multimodal Data. *Journal on Multimodal User Interfaces* 7 (1–2): 29–37.

Partan, Sarah, and Peter Marler. 1999. "Communication Goes Multimodal." *Science* 283 (5406): 1272–1273.

Perniss, Pamela. 2018. "Why We Should Study Multimodal Language." *Frontiers in Psychology* 9: 1109.

Pfau, Roland. 2012. "Manual Communication Systems: Evolution and Variation." In *Sign Language: An International Handbook*, ed. by Roland Pfau, Markus Steinbach, and Bencie Woll, 513–551. Berlin: De Gruyter Mouton.

Płużyczka, Monika. 2018. "The First Hundred Years: A History of Eye Tracking as a Research Method." *Applied Linguistics Papers* 25 (4): 101–116.

Pouw, Wim. 2024. *EnvisionBOX Modules for Social Signal Processing*. https://github.com/WimPouw/envisionBOX_modulesWP

Pouw, Wim, James P. Trujillo, and James A. Dixon. 2020. "The Quantification of Gesture–Speech Synchrony: A Tutorial and Validation of Multimodal Data Acquisition Using Device-Based and Video-Based Motion Tracking." *Behavior Research Methods* 52 (2): 723–740.

Prince, Kilu von, and Sebastian Nordhoff. 2020. "An Empirical Evaluation of Annotation Practices in Corpora from Language Documentation." In *Proceedings of the 12th Conference on Language Resources and Evaluation* (LREC 2020), ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2778–2787. Paris: European Language Resources Association.

Puupponen, Anna, Tuija Wainio, Birgitta Burger, and Tommi Jantunen. 2015. "Head Movements in Finnish Sign Language on the Basis of Motion Capture Data: A Study of the Form and Function of Nods, Nodding, Head Thrusts, and Head Pulls." *Sign Language and Linguistics* 18 (1): 41–89.

Rasenberg, Marlou, Aslı Özyürek, and Mark Dingemanse. 2020. "Alignment in Multimodal Interaction: An Integrative Framework." *Cognitive Science* 44 (11): e12911.

Rau, Felix. This volume. "Language Archives in Context." In *The Documentarist Turn: From Observable Linguistic Behaviour to Typological Generalizations*, ed. by Sonja Riesberg, Uta Reinöhl, and Birgit Hellwig. Amsterdam: John Benjamins.

Ravelli, Louise. 2024. "Centering Multimodality: A Personal and Disciplinary Journey." *Multimodality and Society* 4 (3): 259–274.

Rees, Alice, Lewis Bott, and Petra B. Schumacher. 2019. "Event-Related Potentials in Pragmatic Priming." *Neuroscience Letters* 712: 134435.

Reineke, Silke, Arnulf Deppermann, and Thomas Schmidt. 2023. "Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch (FOLK)." In *Korpora in der germanistischen Sprachwissenschaft: Mündlich, schriftlich, multimedial*, ed. by Arnulf Deppermann, Christian Fandrych, Marc Kupietz, and Thomas Schmidt, 71–102. Berlin: De Gruyter Mouton.

Rochet-Capellan, Amélie, and Susanne Fuchs. 2014. "Take a Breath and Take the Turn: How Breathing Meets Turns in Spontaneous Dialogue." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (1658): 20130399.

Rohrer, Patrick Louis, Ronny Bujok, Lieke van Maastricht, and Hans Rutger Bosker. 2025. "From 'I Dance' to 'She Danced' with a Flick of the Hands: Audiovisual Stress Perception in Spanish." *Psychonomic Bulletin and Review* 32 (5): 2136–2145.

Rossano, Federico, Penelope Brown, and Stephen C. Levinson. 2009. "Gaze, Questioning, and Culture." In *Conversation Analysis: Comparative Perspectives*, ed. by Jack Sidnell, 187–249. Cambridge: Cambridge University Press.

Rühlemann, Christoph, and Mathias Barthel. 2024. "Word Frequency and Cognitive Effort in Turns-at-Talk: Turn Structure Affects Processing Load in Natural Conversation." *Frontiers in Psychology* 15: 1208029.

Rühlemann, Christoph, and Alexander Ptak. 2023. "Reaching Beneath the Tip of the Iceberg: A Guide to the Freiburg Multimodal Interaction Corpus." *Open Linguistics* 9 (1): 20220245.

Rühlemann, Christoph, and James P. Trujillo. 2024. "The Effect of Gesture Expressivity on Emotional Resonance in Storytelling Interaction." *Frontiers in Psychology* 15: 1477263.

Sandler, Wendy. 2022. "Redefining Multimodality." *Frontiers in Communication* 6: 758993.

Sandler, Wendy. 2024. "Speech and Sign: The Whole Human Language." *Theoretical Linguistics* 50 (1–2): 107–124.

Sandler, Wendy, and Diane Lillo-Martin. 2006. *Sign Language and Linguistic Universals*. Cambridge: Cambridge University Press.

Sarvasy, Hannah S., Adam Milton Morgan, Jenny Yu, Victor S. Ferreira, and Shota Momma. 2023. "Cross-Clause Planning in Nungon (Papua New Guinea): Eye-tracking Evidence." *Memory and Cognition* 51 (3): 666–680.

Satyawati, Ni Made Dadi Astini, Ni Made Sumarni, Ketut Kanta, Josefina Safar, Hannah Lutzenberger, Nick Palfreyman, and Connie de Vos. 2025. "The Balinese Homesign Corpus: New Insights into Corpus Development in a Rural Signing Context." Preprint, OSF.

Sauppe, Sebastian, Caroline Andrews, and Elisabeth Norcliffe. 2023. "Experimental Research in Cross-Linguistic Psycholinguistics." In *The Routledge Handbook of Experimental Linguistics*, ed. by Sandrine Zufferey, and Pascal Gygax, 156–172. New York and London: Routledge.

Savčuk, Svetlana O., and Alexandra Makhova. 2021. "Multimodal Russian Corpus and Its Use in Emotional Studies." *Russian Journal of Communication* 13 (1): 42–61.

Savčuk, Svetlana O., Timofej A. Arxangel'skij, Anastasija A. Bonč-Osmolovskaja, Ol'ga V. Donina, Julija N. Kuznecova, Ol'ga N. Ljaševskaja, Boris V. Orexov, and Marija V. Podrjadčikova. 2024. "Nacional'nyj korpus russkogo jazyka 2.0: novye vozmožnosti i perspektivy razvitija" [Russian National Corpus 2.0: New opportunities and development prospects]. *Voprosy jazykoznanija* 2024 (2): 7–34.

Schegloff, Emanuel A. 1982. "Discourse as an Interactional Achievement: Some Uses of 'uh huh' and Other Things that Come between Sentences." In *Analyzing Discourse: Text and Talk*, ed. by Deborah Tannen, 71–93. Washington, D.C.: Georgetown University Press.

Schembri, Adam, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. 2011. *British Sign Language Corpus Project: A Corpus of Digital Video Data of British Sign Language 2008–2011*. First Edition. https://bslcorpusproject.org/

Schulder, Marc, and Thomas Hanke. 2020. *OpenPose in the Public DGS Corpus*. Version 2. Hamburg University.

Schulder, Marc, and Thomas Hanke. 2022. "How to Be FAIR When You CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages." In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, 164–173. Paris: European Language Resources Association.

Seyfeddinipur, Mandana, Pamela Perniss, and Claudia Wegener. This volume. "Taking Language Use Seriously: On the Importance of Multimodality and Interaction in Documentary Practices." In *The Documentarist Turn: From Observable Linguistic Behaviour to Typological Generalizations*, ed. by Sonja Riesberg, Uta Reinöhl, and Birgit Hellwig. Amsterdam: John Benjamins.

Stephenson, Lisa J., S. Gareth Edwards, and Andrew P. Bayliss. 2021. "From Gaze Perception to Social Cognition: The Shared-Attention System." *Perspectives on Psychological Science* 16 (3): 553–576.

Vigliocco, Gabriella, Pamela Perniss, and David Vinson. 2014. "Language as a Multimodal Phenomenon: Implications for Language Learning, Processing and Evolution." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (1651): 20130292.

Vos, Connie de, and Ulrike Zeshan. 2012. "Introduction: Demographic, Sociocultural, and Linguistic Variation across Rural Signing Communities." In *Sign Languages in Village Communities*, ed. by Ulrike Zeshan, and Connie de Vos, 2–24. Berlin: De Gruyter Mouton.

Wiener, Seth, and Kiwako Ito. 2014. "Do Syllable-Specific Tonal Probabilities Guide Lexical Access? Evidence from Mandarin, Shanghai and Cantonese Speakers." *Language, Cognition and Neuroscience* 30 (9): 1048–1060.

Williams, Nikolas S., Genevieve M. McArthur, Bianca de Wit, George Ibrahim, and Nicholas A. Badcock. 2020. "A Validation of Emotiv EPOC Flex Saline for EEG and ERP Research." *PeerJ* 8: e9713.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. "ELAN: A Professional Framework for Multimodality Research." In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, ed. by Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, 1556–1559. Paris: European Language Resources Association.

Yngve, Victor H. 1970. "On Getting a Word in Edgewise." In *Papers from the 6th Regional Meeting of the Chicago Linguistic Society*, 567–577. Chicago, Illinois: University of Chicago.

Zima, Elisabeth. 2017. "On the Multimodality of [all the way from X PREP Y]." *Linguistics Vanguard* 3 (s1): 20160055.

Zima, Elisabeth, Peter Auer, and Christoph Rühlemann. 2025. "Why Research on Gaze in Social Interaction Needs Mobile Eye Tracking." In *Mobile Eye Tracking: New Avenues for the Study of Gaze in Social Interaction*, ed. by Elisabeth Zima, and Anja Stukenbrock, 24–66. Amsterdam: John Benjamins.

CHAPTER 10

Multimodal communication in historical perspective

Gestures and visual prominence in Pharaonic Egyptian monumental inscriptions

Silvia Kutscher
Humboldt Universität zu Berlin

The study of multimodality in communication is a field of research that has been receiving increasing attention in various humanities disciplines since the 1990s. While research on multimodality deals extensively with communicative situations in recent European and Anglo-Saxon cultures, such situations in ancient times have hardly been considered. The article discusses some of the challenges and potentials of multimodality research for the study of historical languages using the example of ancient Egyptian monumental inscriptions. It focuses on two topics. Firstly, the question is examined whether and to what extent historical inscriptions can be used for gesture research. Secondly, the parameter of size scaling is used to show that perceptual saliency and visual prominence are relevant for the constitution of meaning in monumental inscriptions.

Keywords: multimodality, gesture, emblems, performative gesture, semantic macrostructure, information packaging, Ancient Egyptian, monumental inscription, text-image-composition, hieroglyphic script

1. Introduction

Egyptian hieroglyphic writing, whose use has been documented since around 3320 BCE (Kahl 2001; Regulski 2016), presumably represents the oldest known writing system of mankind and was utilised in Ancient Egypt for a period exceeding 3500 years. The system is distinguished by a complex system of different sign function classes,[1] comprising, among others, phonetically interpretable

1. For a detailed description of the hieroglyphic writing system, see Schenkel (1984, 2003, 2011), Kammerzell (1998), Polis and Rosmorduc (2015), and Polis (2018, 2023).

https://doi.org/10.1075/slcs.240.10kut
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

graphemes representing the consonants of the spoken language (1a), as well as those that are of a semantic reading, such as logograms (1b) and classifiers[2] (1c).

(1) a.  *m*

b.  *zwr* 'drink'

c. 

msḏ^{3} CLF 'to bow to someone'

A distinctive characteristic of the Egyptian hieroglyphic writing system is its pictorial nature (Schenkel 2011; Seidlmayer 2012, among others). However, while all hieroglyphs represent objects, humans or animals in a recognisable way, this pictoriality is connected to the meaning of the sign only for those classes which have a semantic reading, such as logograms and classifiers. This is not the case for glyphs representing consonants of the spoken language, as is illustrated in Example (1). The glyph in (1a) is pictorial in nature, as it adequately represents an owl. However, it is important to note that its pictoriality has no correlation to its reading value. In contrast, the logogram in (1b) has a reading value that is also manifested in the pictorial nature of the glyph. The glyph is a logogram representing the meaning of 'drinking'. The image in question depicts a person engaged in the act of drinking using a vessel (Lincke & Kutscher 2012: 128). The pictorial nature of the glyph forms thus establishes an iconic link between the form of the grapheme and its meaning. The same observation pertains to classifiers, as the classifier hieroglyph that is written after the verb *msḏ*^{3} 'to bow to someone' in (1c) illustrates.[3] The classifier represents a human person bending the upper body. What can also be seen is that the pictoriality of the hieroglyphs deliver supplementary information about the performance of daily activities (e.g. that for drinking a small bowl is used) as well as socio-culturally normed gestures. For instance, the classifier in (1c) illustrates that the bowing gesture of reverence is performed not only with a bent back but also with the arms held out in front of the body. In contrast to other writing systems that were also documented at an early stage and used over a longer period of time, such as cuneiform, the hieroglyphic writ-

2. Classifiers (also referred to as "determinatives") exists only in the written language. The hieroglyphic classifiers are written after a word or a phrase and categorize these expressions according to semantic classes (Kammerzell 1999, 2015; Goldwasser 2002; Lincke 2011; Lincke & Kammerzell 2012).

3. Cf. https://thesaurus-linguae-aegyptiae.de/lemma/75660, in: *Thesaurus Linguae Aegyptiae*, (11/06/2025).

ing system of Ancient Egypt retained its pictoriality throughout its entire period of active use.

Another significant feature of the writing system is the possibility of writing texts both from left to right and from right to left. Those graphemes which have an inherent front due to their pictorial nature, e.g. glyphs looking like drawings of bodies of human figures and animals, must be oriented in the same direction. This linear alignment specifies the reading direction of the grapheme sequences, which is opposite to the direction in which the figures and animals are looking. A change in the reading direction within a text is permissible and may be repeated several times within a text. This change in the reading direction can have different functions. Among other things, it can be used for the internal structuring of complex texts or for the layout design of text-image compositions. An example of the application of such a change in reading direction can be found in this article in Section 3.1, Figure 1. There, the writing direction changes between the thematic blocks of the testamentary disposition.[4] Since writing glyphs could be very time-consuming, two cursive forms of hieroglyphic writing developed in parallel, both of which exhibit a decrease in iconicity potential: cursive hieroglyphs and hieratic. The use of each type of writing was largely determined by factors relating to the communicative situation and communicative goals or functions as well as the materiality of the writing medium.[5] Non-cursive Hieroglyphic writing was used in particular for the creation of monumental inscriptions.

The study of the semiotics of Ancient Egyptian hieroglyphic multimodal communication is a fruitful field of research, particularly for two reasons. Firstly, the iconicity of logograms and classifiers plays a decisive role in the genesis of characters (Lincke & Kutscher 2012) and in the interplay of writing and images in the context of multimodal communication acts (Seidlmayer 2012; Baines 2007; Lapčić 2014; Kammerzell 2021; Kutscher to appear). Secondly, many of the earliest examples of Egyptian writing were text-image compositions (Kammerzell 2021). This opens up the possibility of investigating the initial emergence and further development of compositional techniques and layout styles, as well as the formation of multimodal registers and genres in an emerging writing culture. A central method of text-image composition in monumental inscriptions in Ancient Egypt is that of differences in scaling of textual as well as pictorial elements within a multimodal text and the associated visual prominence of elements of the text-image composition. This topic is discussed in Section 3.

---

4. In exceptional cases, individual glyphs can also be rotated within a line, which produces specific semiotic effects concerning the meaning making within the text (Kutscher 2025).

5. For the writing systems used in parallel with hieroglyphics in Pharaonic Egyptian culture, see van Lieven and Lippert (2016).

The monumental architecture of Pharaonic Egypt focuses primarily on sacred and funerary buildings. Residential buildings and palaces, which once served as residences for the elite and the royal family, were not built of stone and are no longer preserved. Sacred buildings are usually decorated with a variety of inscriptions. The majority of the communication acts depicted consist of text-image compositions (Baines 2008: 99). Insofar as the financial resources of the tomb owners allowed, this also applied to their tombs. Monumental inscriptions are usually scenic in nature and represent a variety of socio-culturally established practices and communication situations of the Ancient Egyptian elite. However, it should be noted that these practices are all limited to the religious sphere (e.g. communicating with the gods, burial rituals, providing for the dead through offerings, the afterlife) as well as certain social and legal affairs (biographic events, donations, wills) and political acts (e.g. receiving tribute, acts of war, traveling on behalf of the king). It should also be noted that the principles of representation in Ancient Egyptian images are not based on depicting events as realistically as possible. Naturalistic pictorial representations of spatial relationships, axes of vision, and the physical properties of entities are not relevant to Pharaonic Egyptian iconography. For instance, in Pharaonic Egyptian monumental text-image compositions, it is often the case that one of the pictorial figures is significantly larger in scale than other figures in a scene. In Egyptology, this phenomenon is referred to as "hierarchic scaling".[6] Furthermore, many of the scenes depicted are accompanied by explanatory captions that contain speech reproductions or contextualise the situation depicted. This raises the question of the extent to which historical inscriptions can be used to study (multimodal) text gestures, which represent another area of multimodal communication. Section 4 explores this question. But first, Section 2 explains the theoretical background of multimodality research, which is relevant to the considerations on multimodality in Ancient Egypt presented in Sections 3 and 4. The fifth section summarises this article.

6. This term refers to the fact that the size of the figures depicted encodes their relative social status. That social status differences cannot be seen as the sole explanation for the "hierarchic scaling" but from a multimodal perspective is a relevant means of information packaging, will be discussed in Section 3.2.

2. Multimodality research

The study of multimodal communication is a field of research that has been gaining attention in various humanities disciplines since the 1990s (for an overview, see, among others, Bateman 2008, 2014; Bohle 2014; Diekmannshenke et al. 2011; Jewitt ed. 2014; Kress & van Leeuwen 2001; Klug & Stöckl eds. 2016; Müller et al. 2014; Norris 2015; Wildfeuer et al. 2020). As this is a relatively young field of research, no uniform theoretical model has yet emerged, and the concept of modality is strongly influenced by the individual disciplines involved. Two subbranches of the field, namely gesture research and social-semiotically informed research on multimodality in graphic communication are of relevance for this paper.

2.1 Gesture research

The term gesture is used in many disciplines and thus varies substantially in its meaning (Streeck 2015: 27). In linguistics, gesture is understood as those movements of the hands which are semiotically effective in an act of communication. Gestures in this sense can be performed consciously or subconsciously and accompany speech or replace speech. Within this domain of research, the understanding of what modality is, hence focuses primarily on the sensory conditions of communication, i.e. a modality is a perception channel by which communication is mediated. Multimodality then is a communicative act involving at least two perception channels, such as the auditory channel (i.e. speaking) accompanied by the visual one (i.e. gestures). Human gestures in communication is a field of research that is being investigated in several disciplines. Accordingly, there are numerous proposals for defining and classifying communicative gestures, which unfold along different parameters (for overviews Müller 1998; Bohle 2014; Streeck 2015; Müller et al. 2014). One of the most influential up to this day is the classification of Ekman and Friesen (1969), which will be the basis for this article. Since gestures belong in the domain of bodily behaviour, the first broad differentiation to be drawn is between hand movements that have a communicative function and those that are not meaning related within a discourse. This differentiation was first introduced by Efron [1941] (1972: 96); cf. Bohle (2014: 1454) and later refined by Ekman and Friesen (1969), which termed this class of gestures "adaptors". Adaptors comprise, amongst others, hand movements that are used to manipulate objects such as handling tools and materials.

Within the domain of gestures with communicative functions, the classification of Ekman and Friesen (1969) differentiates between "regulators" of the social interaction within a discourse, and "affect displays" as well as "emblems" and "illustrators". The latter comprise those communicative hand movements which

accompany spoken language and can vary greatly in their individual forms and occurrence. Illustrators are joined with the spoken language and in normal discourse are not used without speech. They normally are executed subconsciously and support or reflect the meaning of the spoken words. They do not have a fixed, easily translatable meaning of their own. In contrast, "emblems" — sometimes also called quotable gestures or symbolic gestures — are gestures, which do have a meaning of their own which is easily translatable into language. They have not only a conventionalised meaning but are also strongly conventionalised in their form and are executed deliberately within a discourse. Emblems may be used outside speech or replace portions of speech in discourse (Ekman & Friesen 1969; Müller 1998; McNeill 2000; Kendon 2004). They have been subclassified in several ways (McNeill 2000; Payrato 2014; Poggi 2014; Teßendorf 2013). A special class relevant for this article consists of conventionalised gestures which are uttered in sociocultural settings of an official character, e.g. discourse in more or less ritualistic settings in which matters are negotiated which are either of religious nature or have social functions and consequences, such as political or legal affairs. The gesture used in those situations has a performative illocutionary force, i.e. the gesture itself is necessary to legitimise the communicative act (e.g. an oath taking gesture). Following Müller (1998), they will be called "performative gestures" ("performative Geste", Müller 1998: 43) in this paper.

2.2 Multimodal graphic communication

Since the subject of investigation is far removed from our present in terms of time and technology, research into linguistic communication in antiquity is based solely on visual notations of language in writing. However, written texts are more than just sequences of graphemic representations of spoken language elements. They are graphic objects whose visual aspects (e.g. graphetics, layout) are just as much involved in the constitution of meaning as the linguistic information conveyed by the graphemes. In German-language research, this has already been established by Ehlich (1984: 23) and Raible (1991). Since Barthes (1964) at the latest, it has also been known that written language, the visual design of the graphic surface, and images form a communicative unit in a synergistic interaction. The investigation of these aspects of graphic communication established itself as an independent field of research, in particular through Kress and van Leeuwen (1996, 2001, 2006). They represent the starting point for semiotically oriented multimodality research and show that investigating this multimodal interaction provides a necessary new perspective for text linguistics in understanding how meaning is constituted in texts. Accordingly, within this domain of research, the understanding of what modality is, takes a differ stance from linguistic gesture research traditions. Anchored in Sys-

temic Functional Linguistics (Halliday 1978, 1985), Kress and van Leeuwen (1996, 2001, 2006) assume that the constitution of meaning is the result of social action and therefore modes are defined as "socially shared and culturally given resources for making meaning" (Kress 2009: 54). The meaning of a communicative act is thus seen as the result of the complex interplay of various resources perceived by the communication partners as semiotically effective, regardless of the perception channel they are produced or perceived in (Bateman et al. 2017). These semiotic resources include, on the one hand, culturally established sign systems such as language, writing, and images. On the other hand, there are sociocultural and other situational factors. These include the social constellations between the communication partners, the materiality of the communicative act, and its spatial conditions (Bateman 2016, 2021; Wildfeuer & Bateman 2018; Kutscher 2025; Kutscher & Serova 2025). Multimodality research has recently branched out into a number of different approaches and further developments of older models. Good overviews are provided by handbooks and introductions to multimodality research, including Bateman (2014), Bateman et al. (2017), Jewitt (2009, ed. 2014), Jewitt et al. (2016) and Klug and Stöckl (eds. 2016).

While research on multimodality has focused intensively on communication situations in recent European and Anglo-Saxon cultures, those of antiquity are only gradually finding their way into the multimodality discourse, in Egyptology, e.g. Lincke and Kutscher (2012), Lapčić (2014), Döhl et al. (2020), Kutscher (2020), Kammerzell (2021), Jurman (2023), Rogner (2022), and all articles in Kutscher and Serova (eds. 2025).

The following aspects of research on multimodality in graphic communication are of particular relevance for the ideas presented in this article. Firstly, the aspect of dimensionality. As material objects, (multimodal) texts have both a linear and a two-dimensional nature. Since holistic processing of the represented content is not possible for cognitive reasons, the recipient must decide on an entry point into the process of meaning constitution for the reception process as well as its further sequencing. Writing systems have standardised indices that indicate a linear orientation on the horizontal plane, thus signalling the entry point intended by the producer and the intended reading sequence within lines of text. Graphic communication objects are usually also pre-structured with regard to their vertical axis by sociocultural norms or praxeological knowledge. These rules can vary depending on the medium or genre. While the entry point and further navigation for reading is standardised in the medium of a book — for example, one usually starts at the top left of the first page of the first chapter and end on the bottom of the last page to ensure a coherent process of meaning making — this does not apply e.g. to a newspaper page with several articles. Here, the entry point and reading sequence is not standardised; different entry points are possi-

ble because readers can choose different articles, for example. The producer can attempt to control the reader's reception process by using different graphic means to set sequencing signals for the intended entry point and the reading order of the individual text zones (cf. "reading path", Royce 1998; "navigation", Bateman 2008; Cohn 2013). An important means of doing this is the higher perceptual salience of large-scale elements compared to smaller ones. From a perceptual psychology perspective, differences in the relative size of graphic elements are weighted differently in perception.[7] The visually prominent element — i.e. the largest of the elements within a graphic zone — also acts as an eye-catcher and marks the point of entry into the reception process intended by the producer. In the case of a newspaper page, for example, this would be a large headline.

This semiotic effect of scaling leads us to the second aspect relevant for this article, namely the weighting of information within the semantic macrostructure of the multimodal text. As numerous works in text linguistics outside the multimodality paradigm have already shown, meaning constitution in the reception of texts takes place on two levels: (1) the local text level, where the cohesion and coherence relationships between successive sentences are examined, and (2) the global text level, i.e. the entirety of the text (Hellwig 1984; Raible 1978; van Dijk 1980; von Stutterheim & Klein 1989; for a recent overview of further approaches developed later, see also Heine 2019). At this global text level, local information is interpreted in terms of an overall meaning and a global topic is inferred ("topic of discourse", van Dijk 1980: 41). In the reception process, information is therefore not only linked locally — i.e. from sentence to sentence — but also weighted at the global level in terms of its communicative relevance for the entire text, cf. "semantic macrostructure" in van Dijk (1980: passim). For example, a "discourse referent" (van Dijk 1980: 41, cf. also Karttunen 1969: 1) can be identified as the central protagonist of the global plot and thus interpreted as the major discourse referent. These text-linguistic findings can be transferred to the analysis of the visual design of multimodal graphic objects (Kutscher 2020), investigating the different scaling of graphic zones (Kutscher 2020, to appear) is an important graphic means for the weighting of information within the meaning making process. Due to cognitive effects of perception (cf. Gestalt psychology, e.g. Wertheimer 1923), larger scaled zones of a graphic object are more salient in perception and are processed semiotically before smaller scaled elements. The graphic device of size scaling thus contributes to guiding the viewer's gaze and controls the sequencing of the reception process (Royce 1998, 2007; Kress & van Leeuwen 1996, 2006; Bateman 2008; Cohn 2013, among others). Elements with the highest degree of scaling within a

7. See also Kress and van Leeuwen (1996, 2006), who refer to Arnheim (1954) and also to Gestalt psychology.

graphic object are also visually prominent. They highlight the largest element at the level of the semantic macrostructure of the (multimodal) text. Visual prominence signals that the highest scaled element is of particular relevance for the constitution of meaning. It thus contributes significantly to the construction of the recipient's semantic macrostructure. The graphic device of size scaling therefore has two semiotic effects within the reception process: It guides the recipient's gaze (among others, Royce 1998, 2007; Kress & van Leeuwen 1996, 2006; Bateman 2008; Cohn 2013) and at the same time serves to control the packaging of information at the macrostructural level of the multimodal text (Kutscher 2020).

3. Investigating the semiotic effects of scaling in Pharaonic Egyptian monumental inscriptions

While scaling as graphic means for structuring the reading process as well as the meaning constitution of a text has been attended to in multimodality-oriented research, studies on Pharaonic Egyptian inscriptions has so far hardly been conducted from this perspective. In the following, both aspects of the scaling of graphic zones within a multimodal text will be illustrated using a hieroglyphic Egyptian inscription from the tomb of Wep-em-nefert in Giza (G 8882) from the 5th Dynasty of the Old Kingdom (ca. 2440 BCE), see also Kutscher (2020, to appear).[8]

3.1 Guiding the recipient's gaze by scaling

The multimodal text used for the following considerations comes from a tomb complex consisting of two graves and two burial chapels. The initiator and owner of the tomb complex is the high official Wep-em-nefert, who bequeathed the second burial chapel and the second burial shaft of the complex to his eldest son Iby. Wep-em-nefert's testamentary disposition was engraved on the east wall of the chapel for this second tomb. The complete contents of the will are presented in Figure 1. The text is a complex composition of text and image elements, with the text elements written in hieroglyphic script. The glyphs in the various textual zones are oriented either from right to left or from left to right. The various graphic zones also vary in scale, see (1).

8. Full documentation of the tomb, including translation of inscriptions, is given in Hassan (1936, vol. 2: 179–201 and plates LXIV–LXXVI). Translations are also given by the *Thesaurus Linguae Aegyptiae*: https://thesaurus-linguae-aegyptiae.de/object/QL6PRYR7JRG6BCLFM2V4LQDUDY, in: *Thesaurus Linguae Aegyptiae,* (11/06/2025).

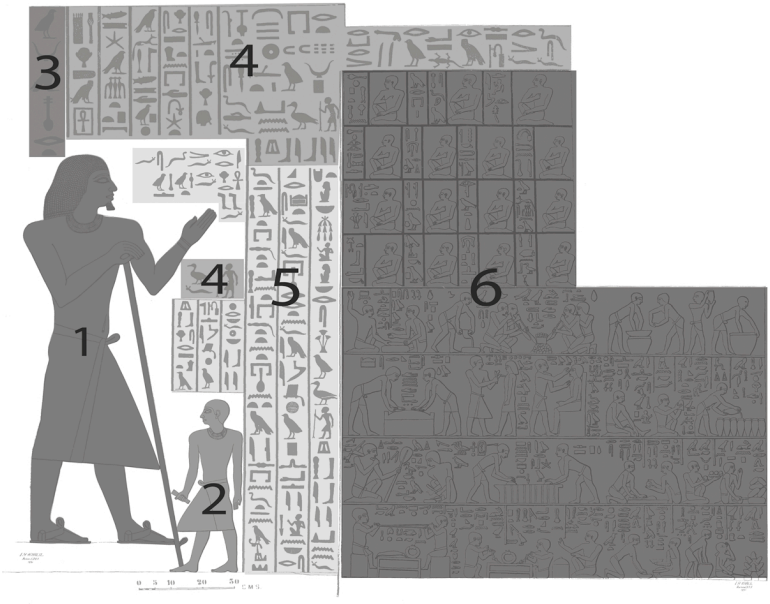


Figure 1. Different scaling grades (1–6) in the relief concerning the will of Wep-em-nefert in G8882. Line drawings of relief from Hassan (1936: 191)

The relief shown in Figure 1 depicts a series of text zones of varying scales (scales 3–6). In addition, some text zones are aligned to the left, while others are aligned to the right.

The largest text zone (size 3) in Figure 1 identifies the large figure (size 1) as the high-ranking official named Wep-em-nefert. From the architectural surroundings and the sociocultural context, he can be clearly identified as the owner of the tomb in which the relief is located. The name column serves as an entry point for understanding the textual elements of the relief. This is the declaration of Wep-em-nefert's last will and testament, in which he bequeaths part of the tomb complex to his eldest son (Hassan 1936; contains the complete transcription and translation of the relief). The latter is depicted in size 2 standing at his father's feet.[9] The text passages written in scale size 4 with the same orientation in the header of the relief describe Wep-em-nefert in more detail as a royal official by naming his titles. These provide information about his social status and his social duties. In the same scale, but with the orientation reversed, the type of document, its date of issue, the heir, and the fact that the persons listed in the upper right part

9. The isolated text section in size 4 reads "his eldest son" and, by using the same font size, graphically establishes a relationship to the heading of the will. The phrase refers to the heir already mentioned in the heading of the will.

of the image (size 6) were present as witnesses when the legal text was issued are mentioned.[10] The lower right image field (size 6) shows the production of grave goods. The craftsmen depicted are shown in dialogical scenes, so-called "daily-life scenes", producing grave goods.

The scaling of the text units initially draws the eye to the name of the tomb's owner (size 3). The text scaled to size 4 provides information about the type of text and the social status of its initiator. The content of the will is scaled smaller (size 5). The scaling of the text units thus suggests a semiotically efficient reading order that maximally supports the reception process. This design of the relief in Figure 1 is not unusual or exceptional for monumental inscriptions. Rather, it has been the standard practice in the design of monumental reliefs since the Old Kingdom (i.e. from approx. 2670 BCE onwards). It can therefore be assumed that the producer of the text intended this sequence of information processing and that the scaling was therefore deliberately chosen.[11]

3.2 Weighting of information by visual prominence

The semiotic effect of visual prominence on the processing of the semantic macrostructure can be illustrated by considering the figurative representation of the tomb owner Wep-em-nefert in Figure 1. As already explained, the large figure on the left of the relief represents the initiator of the tomb complex in which the relief is located. In relation to the message of the composition, Wep-em-nefert represents the major discourse referent of the multimodal text. He is the client who commissioned the construction of the tomb and made the testamentary disposition that is reproduced in the relief. As the testator, he determines the estate and the heir. The assessment of Wep-em-nefert as central in the semantic macrostructure of the composition is also supported by the length of the reference chain: the number and spatial distribution of expressions that refer to Wep-em-nefert is significantly greater than that of expressions that refer to other referents in the multimodal text, for details, see Kutscher (2020).

---

10. Text of header, as translated by Hassan (1936:190), right-to-left orientation: "Wep-em-nefert. The sole confidant, the chief *nekhbite*, the master of the secrets of the toilet house, the priest of Horus and Anubis, the domain administrator of Buto, the domain administrator of [the vineyard called] "Star-of-Horus" president of heaven, the director of the palace and the director of the distribution of food in the house of life." Left-to-right orientation: "The year of joining the two lands, third month of the season of winter, day 29th. The sole confidant Wep, he says:"

11. In addition to scaling, other graphic design elements should also be considered in the analysis. A detailed examination of these elements and their semiotic effects would go beyond the scope of this article. Kutscher (2020) provides a detailed analysis of the reception process controlled by graphic elements.

In Egyptology, this phenomenon is also referred to as hierarchic proportion, hierarchic scaling or *Bedeutungsmaßstab*. This term reflects the assumption that the size of the figures depicted encodes a difference in power relations or social status. However, social status differences cannot be seen as the sole explanation for the hierarchic scaling. For example, figurative representations of wives of socially high-ranking individuals are sometimes significantly smaller than the figures of their husbands, and sometimes of equal size, which is problematic from this standpoint since the social rank of wives are the same in each case (Vasiljević 1998; Roth 2006). Therefore, the function of size scaling of figures within the same scene or scene groups has been subject to some debate, e.g. Vasiljević (1998), Roth (2006). Several motivations for the function of hierarchic scaling have been given, but have not been found to sufficiently explain the complex phenomenon (differences in scaling sizes, posture and contexts of occurrence differ across times and places; cf. Roth 2006 for an overview in the Old Kingdom period). Both Vasiljević (1998) as well as Roth (2006) show that differences in social status are not convincing explanations for the scaling differences and argue for taking the context of scenes into consideration. However, not all configurations exhibiting hierarchical scaling can sufficiently be explained in the ways suggested by Vasiljević (1998) or Roth (2006). The text-linguistically informed approach presented in this paper — namely to understand it primarily as a graphic information packaging strategy that operates at the semantic macrostructure of the multimodal text — seems to offer a new perspective which can be applied to all types of configurations which show hierarchic scaling. This working hypothesis, however, needs a deeper investigation into a larger amount of adequate data.[12]

12. The investigation of large amounts of data additionally are best supported through visualizations of the analysis. Consequently, data corpora should be designed in such a way that they can visually represent complex semiotic relationships within graphic communicates and enable multimodal visual annotation. The project *CaeMmCom — Corpus of Ancient Egyptian Multimodal Communication* (https://www.archaeologie.hu-berlin.de/de/aknoa/forschung-und-projekte/projekte/caemmcom, 11/06/2025), based at Humboldt University in Berlin, is working on the conception and implementation of a digital research environment that makes this possible. The basis for the visualisation of the data in the corpus is the open-source software *HyperImage* and its further development *Yenda1*. The architecture of the corpus is largely based on the parameters illustrated in Kutscher (2020). A more detailed description of the structure of the corpus can be found in Döhl et al. (2020).

4. Investigating gestures in Pharaonic Egyptian monumental inscriptions

It is evident that video recordings of everyday communication in ancient language communities cannot exist. Statements about spontaneous gestures that accompany speech in social interaction (i.e. "illustrators" in the sense of Ekman & Friesen 1969) cannot therefore be made for Ancient Egypt. There are pictorial representations of dialogical scenes in the text-image genre known as "daily-life scenes" in Egyptology. In this genre, texts that imitate spoken language are composed with pictorial figures to create scenes that illustrate various craft activities. The scenes show figures conversing with other craftsmen or supervisors during the work process. However, the hand movements depicted are determined by the work process shown, i.e. they have no communicative function. They can be classified as "adaptors" in the sense of Ekman and Friesen (1969). For this reason, they do not allow any conclusions to be drawn about the form and function of spontaneous communicative gestures that accompany speech.[13]

Notwithstanding the existence of scenes depicting "daily life", there are various pictorial representations of human figures in social interaction which do not involve the handling of objects or materials. In these depictions, various socio-culturally established situations are illustrated. Contents of the scenes as well as the design of the depictions follows specific iconographic principles.[14] They are highly conventionalised in form as well as in function. Evidence coming from text elements accompanying these scenes suggests that the gestures illustrated in the scenes are of a communicative nature. This assertion is supported by archaeological findings, which facilitate a more comprehensive understanding of the depicted social situations in the images.[15] By analysing the pictorial representations of hand movements in these depictions it is possible to examine gestural behaviour in different socio-cultural normed situations. Since the represented gestures occur not only regularly in these depictions but gestures occurring in the same social situations are also similar in form,[16] it can be assumed that these are

13. In (1), such scenes can be seen in the lower half of the area marked with the number 6.

14. For ideologically-driven constraints on social interaction and the aesthetics of pictorial and textual representation in Ancient Egyptian culture, see Baines (2023).

15. Complementing these data, knowledge can also be increased by investigating linguistic formulations in texts as well as the iconicity of hieroglyphic graphemes (Dominicus 1994), e.g. the classifier hieroglyph illustrated in (1c) in Section 1.

16. Apart from stylistic differences.

conventionalised gestures that were most likely obligatory in the respective communication acts.[17]

However, categorizing these conventionalised gestures poses a challenge for analysis. The contextualising textual elements that accompany the scenes do not always provide information about whether they served to accompany speech or were performed without speech. For example, in the case of gestures executed by socially inferior figures interacting with higher-ranking figures it cannot be determined with certainty whether the gestures of address or greeting were performed to accompany a greeting formula or whether they functioned as emblematic, potentially speech-replacing gestures of deference.

The accompanying text for the gesture of the large figure of Wep-em-nefert in (1) in Section 3.2, on the other hand, is introduced with the predicate *ḏd-f* 'he speaks'. The gesture of the bent arm with the open hand stretched upwards is regularly found in scenes depicting communicative situations of official speech, such as legal and religious discourses and speech in a royal context. The gesture is a conventional gesture accompanying a legally relevant speech taking place in a public space. However, based on the available data, it cannot be clearly determined whether it merely indicates the specific communicative situation — and thus performs a register-identifying function — or whether it performatively legitimises it, i.e. whether it is a performative gesture in the sense of Müller (1998, see Section 2).

The transformation of natural gestures into the pictorial medium requires a reduction of the originally dynamic gestures to a static, two-dimensional form of representation. This can sometimes lead to visual ambiguity with regard to individual gestures. Depending on the underlying theoretical assumptions, these gestures can be interpreted as polysemic or "homonymous", see Figures 2, 3 and 4.

Figure 2 depicts a scene in which two craftsmen are working on the construction of a bed frame. The third craftsman on the far right is identified in the accompanying text as a funerary priest who also is a carpenter.[18] According to his title, the figure hence is the supervisor of the craftsmen working on the bedframe and it may well be assumed that he is pointing to a spot on the bed frame that needs further work, cf. also Dominicus (1994: 155) for the same interpretation.

Figure 3 depicts a herd of cattle being driven across the Nile. A crocodile lurks in the Nile water on the right, threatening to attack the herd. Based on the different contexts in which they are used, Dominicus (1994: chapters III and IV)

17. Supported is this by the fact that the pictoriality of those hieroglyphs that reflect gestures is standardized.

18. Text and translation in *Thesaurus Linguae Aegyptiae*, https://thesaurus-linguae-aegyptiae.de/sentence/IBUBdogkTQcANooqgHcCDHoKaiI, (09/08/2025).

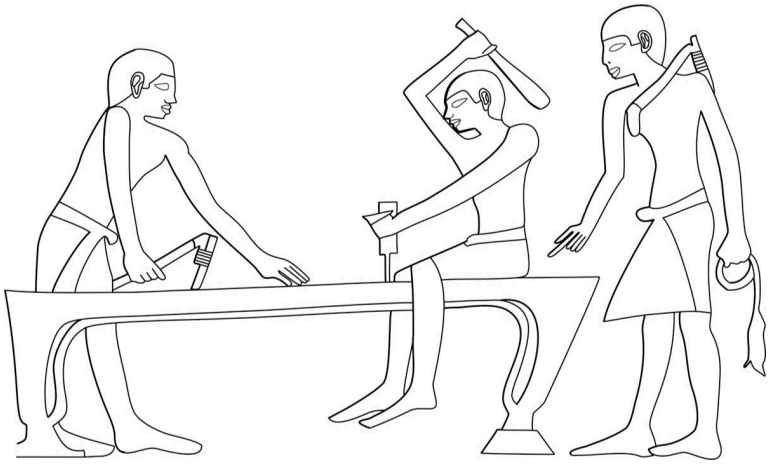


Figure 2. Daily-life scene from the tomb of Nianchchnum und Chnumhotep, representing a supervisor pointing to a spot that needs reworking by the carpenters. Line drawings by O. Tekin after photo in Moussa & Altenmüller (1977, plate 62)

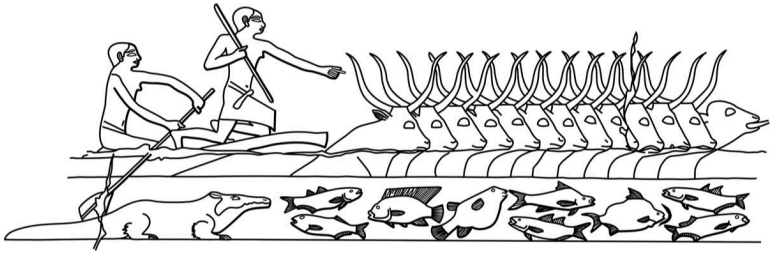


Figure 3. Excerpt from a "daily-life scene" from the tomb of Idut, showing a herd of cattle being driven across the Nile while a crocodile is lurking under the water. Line drawings by O. Tekin after Macramallah (1935, plate VII)

assumes that two different gestures are present: a pointing gesture in (2) and an incantatory gesture in (3). However, this ignores the fact that the hand and arm shapes depicted are the same in both communication situations: an outstretched arm with an extended index finger and a closed hand. Furthermore, it should be noted that the text element accompanying the figure does not contain an incantation. Instead, it is a direct address to the drivers of the herd, which can be translated as a warning to be attentive to the lurking crocodile.[19] A similar phenomenon

19. The translation of the accompanying text of the pointing figure as given by Macramallah (1935: 16) is: "O toi berger, prends garde que ta face soit vivante à ce SY, celui qui est sur l'eau, qui vient sans etre vu; prends bien garde à lui!".

can be observed in other scenes that apparently depict the incantation gesture. In none of the scenes known to date is an incantation spell integrated as an accompanying text element (Vandier 1969: 104–105).[20] Therefore, it seems more plausible to assume that the same gesture is used in both different communication situations, each performing an indexical function. In communication situations in which something is pointed out (2), this indexical gesture draws attention to an entity. In evocation situations (3), the indexical gesture either draws attention to the target of the evocation — in (3), the herd of cattle — , or to the place form where the danger is to be expected — in (3) the waters of the Nile, where the crocodiles are expected to lurk. Instead of two different gestures, we are therefore dealing with only one: an indexical gesture that takes on different meanings in the construction with the linguistic utterance and the situational context.

The case of the multifunctional gesture "arms raised above the head" is somewhat different, cf. Figures 4 and 5. This gesture occurs in two very different contexts, as the respective textual comments on the pictorial scene clearly indicate. In Figure 4, a human figure is shown taken from a scene which represents the funeral of the tomb owner.

The uniformity of the execution and the fact that this type of scene occurs frequently and is always executed in the same way suggest that this is an emblematic gesture of mourning. In Figure 5, however, the same type of gesture is used in the context of jubilation.

The figure depicted in Figure 5 represents a high-ranking official who is being honoured with gold as a reward.[21] Through the gesture depicted, the figure shows his gratitude and joy toward the pharaoh who is bestowing the honour. In this case, due to the strong differences in illocution, it seems more plausible to assume that they are "homonymous", i.e. are two different types of gestures although similar in form. These may as well differ in their natural execution, but they are iconographically similar due to the medium, which only allows static representations of the gestures.

20. "(…) sans doute a-t-il pressenti la présence du crocodile et recommande-t-il à ses compangons le silence et l'attention. Cette hypothèse est confirmé par un tableau de la tombe de Ptahhotep: on voit, en effet, un personage qui fait la même géste de la main droite et qui recommande á une de ses compagnons, qu'il traite de "brenneux", de faire attention au crocodile." (Vandier 1969: 104).

21. For more details of this kind of scene, cf. Baines (2023: 86–87).

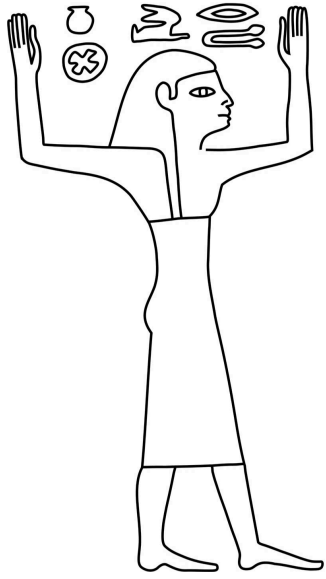


Figure 4. Excerpt of a funeral scene from the tomb of Antefoker, representing a figure in mourning for Antefoker. Line drawings by O. Tekin after Davies (1920, plate 19)

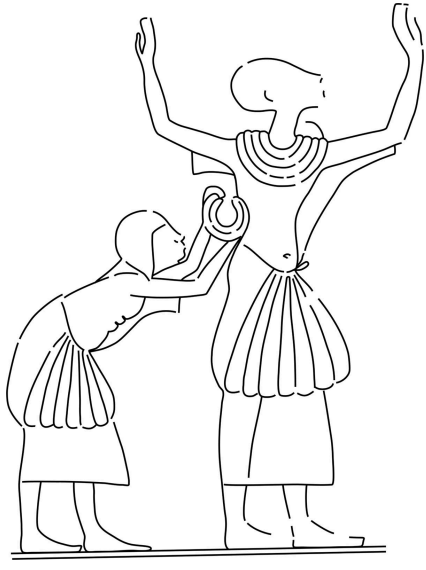


Figure 5. Excerpt of a scene from the tomb of Tutu, representing the tomb owner (right) who is bestowed with gold of honour by the pharaoh (who is placed at a window high above Tutu's head). The person on the left is an official handing over the gold.
Line drawings by O. Tekin after Davies and Ricci (1903, plate 20), https://ia601307.us.archive.org/25/items/cu31924020525394/cu31924020525394.pdf, (11/6/2025)

5. Conclusion

While research on multimodality has focused extensively on communicative situations in modern European and Anglo-Saxon cultures, such situations in ancient times have received comparatively little attention. The objective of this study was to utilise monumental inscriptions from ancient Egyptian eras to illustrate the potential for multimodality research to be successfully employed in the analysis of historical cultural writings. In particular, two subbranches of multimodality research have been discussed: namely, linguistic gesture research and research on multimodality in graphic communication. The paper sets out the challenges and potential of gesture research for the study of historical languages, based on theoretically informed research on gestures in modern spoken languages. It has been posited that, despite the impenetrability of the spoken-linguistic use of gestures, the textual elements accompanying the representations and archaeological findings that classify the communicative context of the depicted scenes can be utilised to a certain extent to compensate for this. It has also been demonstrated that adopting a theoretically oriented perspective from linguistic gesture research — considering the form of gestures and distinguishing between conventionalised gestures indicating communicative registers, emblems and performative gestures — can lead to a deeper understanding of the inventory and function of conventionalised gestures in Ancient Egypt. Moreover, the paper has demonstrated that the multimodal approach can also be applied fruitfully to analyse the layout of Ancient Egyptian monumental inscriptions. In order to comprehend the meaning-making process of a communicative act in as thorough a manner as possible, it is necessary to analyse semiotic processes in several steps. One of these steps involves the analysis of the graphic means used to guide the reception process in a manner intended by the producer of the inscription. To illustrate this phenomenon, the parameter of size scaling was utilised. It has been posited that scaling, encompassing both various scaling grades for navigational purposes and the visual prominence of the element of the highest scaling grade, functions as a significant graphic device employed in multimodal monumental inscriptions. It was argued that this device has been deliberately implemented by the producers of these inscriptions in ancient Egypt to manage the meaning-making process of the recipient. Given that hieroglyphic writing is the origin of writing and was in use for a period of over three thousand years, research on the multimodality of graphic communication in Ancient Egypt not only profits from the theory of multimodality research. Furthermore, it may also contribute substantially to the development of theory, especially in the field of the genesis and development of layout design, communicative registers and (multimodal) text genre.

Funding

This article was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) — SFB 1412 Register — 416591334.

Acknowledgments

I would like to thank two anonymous reviewers for their valuable comments.

Abbreviations

Glossing follows the Leipzig Glossing Rules.

References

Arnheim, Rudolf. 1954. *Art and Visual Perception: A Psychology of the Creative Eye*, revised ed. Berkeley: University of California.

Baines, John. 2007. *Visual and Written Culture in Ancient Egypt*. Oxford: University Press.

Baines, John. 2008. "On Functions of Writing in Ancient Egyptian Pictorial Representations." In *Iconography without Texts*, ed. by Paul Taylor, 95–126. London/Turin: Warburg Institute.

Baines, John. 2023. "Ancient Egyptian Decorum: Demarcating and Presenting Social Action." In *(Re)constructing Ancient Egyptian Society: Challenging Assumptions, Exploring Approaches*, ed. by Danielle Candelora, Nadia Ben-Marzouk, and Kathly M. Cooney, 74–89. London: Routledge.

Barthes, Roland. 1964. "Rhétorique de l'Image." *Communications* 4: 40–51.

Bateman, John A. 2008. *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. London: Palgrave MacMillan.

Bateman, John A. 2014. *Text and Image: A Critical Introduction to the Visual/Verbal Divide*. London and New York: Routledge.

Bateman, John A. 2016. "Methodological and Theoretical Issues in Multimodality." In *Handbuch Sprache im Multimodalen Kontext*, ed. by Nina-Maria Klug, and Hartmut Stöckl, 36–74. Berlin: de Gruyter.

Bateman, John A. 2021. "Dimensions of Materiality: Towards an External Language of Description for Empirical Multimodality Research." In *Empirical Multimodality Research: Methods, Evaluations, Implications*, ed. by Jana Pflaeging, Janina Wildfeuer, and John A. Bateman, 35–63. Berlin/New York: de Gruyter.

Bateman, John A., Janina Wildfeuer, and Tuomo Hiippala. 2017. *Multimodality: Foundations, Research and Analysis*. Berlin/Boston: de Gruyter.

Bohle, Ulrike. 2014. "Contemporary Classification Systems." In *Body — Language — Communication*, HSK 38.2, ed. by Cornelia Müller, Alan Cienki, Ellen Fricke, Silvia H. Ladewig, David McNeill, and Jana Bressem, 1453–1461. Berlin: de Gruyter.

Cohn, Neil. 2013. "Navigating Comics: An Empirical and Theoretical Approach to Strategies of Reading Comic Page Layout." *Frontiers in Cognitive Science* 4: 1–15.

Davies, Norman de Garis. 1920. *The Tomb of Antefoker, Vizier of Sesostris I, and his Wife, Senet*. London: Allen & Unwin.

Davies, Norman de Garis, and Seymour de Ricci. 1903. *The Rock Tombs of el-Armana, vol. VI: Parennefer. Tutu, and Aij*. London/Boston: Egypt Exploration Fund, https://ia601307.us.archive.org/25/items/cu31924020525394/cu31924020525394.pdf, (11/06/2025).

Diekmannshenke, Hajo, Michael Klemm, and Hartmut Stöckl (eds). 2011. *Bildlinguistik: Theorien — Methoden — Fallbeispiele*. Berlin: Schmidt.

van Dijk, Teun A. 1980. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*, Hillsdale, NJ: Erlbaum.

Döhl, Rebecca, Silvia Kutscher, and Jens-Martin Loebel. 2020. "CaeMmCom — Corpus Altaegyptischer Multimodaler Communication. Der Aufbau einer Multimodalen Datensammlung Altägyptischer Kommunikate." *Zeitschrift für Digitale Geisteswissenschaften* 5, Version 2.0: 12th May 2023.

Dominicus, Brigitte. 1994. *Gesten und Gebärden in Darstellungen des Alten und Mittleren Reiches*. Heidelberg: Orientverlag.

Efron, David. [1941] 1972. *Gesture, Race and Culture*. The Hague/Paris: Mouton.

Ehlich, Konrad. 1984. "Zum Textbegriff." In *Texte — Textsorten — Semantik: Linguistische Modelle und Maschinelle Verfahren*, ed. by Annely Rothkegel, and Barbara Sandig, 9–25. Hamburg: Buske.

Ekman, Paul, and Wallace V. Friesen. 1969. "The Repertoire of Nonverbal Behavior: Categories, Origins, Usage and Coding." *Semiotica* 1 (1): 49–98.

Goldwasser, Orly. 2002. *Prophets, Lovers and Giraffes: Wor(l)d Classification in Ancient Egypt*. Wiesbaden: Harrassowitz.

Halliday, Michael A.K. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London: Arnold.

Halliday, Michael A.K. 1985. *An Introduction to Functional Grammar*. London: Arnold.

Hassan, Selim. 1936. *Excavations at Giza 2: 1930–1931*. Kairo: Government Press, Bulâq, https://gizamedia.rc.fas.harvard.edu/images/MFA-images/Giza/GizaImage/full/library/hassan_giza_2.pdf. (30/05/2025).

Heine, Bernd. 2019. "Some Observations on the Dualistic Nature of Discourse Processing." *Folia Linguistica* 53 (2): 411–442.

Hellwig, Peter. 1984. "Grundzüge einer Theorie des Textzusammenhangs." In *Texte — Textsorten — Semantik. Linguistische Modelle und Maschinelle Verfahren*, ed. by Annely Rothkegel, and Barbara Sandig, 51–79. Hamburg: Buske.

Jewitt, Carey. 2009. "An Introduction to Multimodality." In *The Routledge Handbook of Multimodal Analysis*, ed. by Carey Jewitt, 1–15. London/New York: Routledge.

Jewitt, Carey (ed). 2014. *The Routledge Handbook of Multimodal Analysis*. London/New York: Routledge.

Jewitt, Carey, Jeff Bezemer, and Kay O'Halloran. 2016. *Introducing Multimodality*. London/New York: Routledge.

Jurman, Claus. 2023. "Non, Je Ne Regrette Rien! A Narratological Essay on Life Writing, Life Telling, and Life Reading in Old Kingdom Egypt." In *Narrative and Narrativity in Ancient Egypt: Case Studies on Narrative Difference in Various Media*, ed. by Gerald Moers, 71–216. Hamburg: Widmaier.

Kahl, Jochem. 2001. "Hieroglyphic Writing During the Fourth Millennium BC: An Analysis of Systems." *Archéo-Nil* 11: 101–134.

Kammerzell, Frank. 1998. "The Sounds of a Dead Language. Reconstructing Egyptian Phonology." *Göttinger Beiträge zur Sprachwissenschaft* 1: 21–41.

Kammerzell, Frank. 1999. "Klassifikatoren und Kategorienbildung in der Ägyptischen Hieroglyphenschrift." *Spektrum. Informationen aus Forschung und Lehre* 3: 29–34.

Kammerzell, Frank. 2015. "Egyptian Verb Classifiers." In *Proceedings of the Xth International Congress of Egyptologists, Rhodes May 22–29, 2008*, ed. by Panagiotis Kousoulis, and Nikolaos Lazaridis, 1395–1416. Leuven: Peeters.

Kammerzell, Frank. 2021. "Reading Multimodal Compositions from Early Dynastic Egypt (With an Appendix on Previously Unlisted, Reinterpreted or Otherwise Noteworthy Signs)." In *Keep out! Early Dynastic and Old Kingdom Cylinder Seals and Sealings in Context*, ed. by Eva-Maria Engel, Annette I. Blöbaum, and Frank Kammerzell, 1–98. Wiesbaden: Harrassowitz.

Karttunen, Lauri. 1969. "Discourse Referents." In COLING '69: Proceedings of the 1969 Conference on Computational Linguistics, *September 1969*, 1–38.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: University Press.

Klug, Nina-Maria, and Hartmut Stöckl (eds). 2016. *Handbuch Sprache im Multimodalen Kontext*. Berlin: de Gruyter.

Kress, Gunther. 2009. "What is Mode?" In *The Routledge Handbook of Multimodal Analysis*, ed. by Carey Jewitt, 54–67. London/New York: Routledge.

Kress, Gunther, and Theo van Leeuwen. 1996. *Reading Images: The Grammar of Visual Design*. London/New York: Routledge

Kress, Gunther, and Theo van Leeuwen. 2001. *Multimodal Discourse. The Modes and Media of Contemporary Communication*. London/New York: Routledge

Kress, Gunther, and Theo van Leeuwen. 2006. *Reading Images. The Grammar of Visual Design*. London/New York: Routledge.

Kutscher, Silvia. 2020. "Multimodale Graphische Kommunikation im Pharaonischen Ägypten: Entwurf einer Analysemethode." *Lingua Aegyptia* 28: 81–116.

Kutscher, Silvia. 2025. "Space as a Semiotic Resource in Stationary Graphic Communication of Ancient Egyptian: A Theoretical Outlook." In *Spaces and Meaning: Multimodal Communication in Ancient Egypt*, ed. by Silvia Kutscher, and Dina Serova, 19–36. Leiden: Sidestone Press.

Kutscher, Silvia. To Appear. "Analysing Ancient Egyptian Text-Image-Compositions: An Outline." In *Multimodal Artefact Analysis in Ancient Studies*, ed. by Patrizia Heindl, and Eva Roßberger. Berlin: de Gruyter.

Kutscher, Silvia, and Dina Serova. 2025. "Spaces and Meaning: Introduction to the Volume." In *Spaces and Meaning: Multimodal Communication in Ancient Egypt*, ed. by Silvia Kutscher, and Dina Serova, 9–18. Leiden: Sidestone Press.

Kutscher, Silvia, and Dina Serova (eds). 2025. *Spaces and Meaning: Multimodal Communication in Ancient Egypt*. Leiden: Sidestone Press.

Lapčić, Aleksandra. 2014. "Bild-Schrift-Gestalten des Göttlichen: Multimodale Informationsverarbeitung im Amduat Thutmosis' III." In *Bild: Ästhetik — Medium — Kommunikation. Beiträge des Dritten Münchner Arbeitskreises Junge Ägyptologie (MAJA 3), 7. bis 9.12.2012*, ed. by Gregor Neunert, Alexandra Verbovsek, and Katrin Gabler, 169–192. Wiesbaden: Harrassowitz.

van Lieven, Alexandra, and Sandra L. Lippert. 2016. "Egyptian (3000 BCE to ca. 400 CE)." In *Biscriptality: A Sociolinguistic Typology*, ed. by Daniel Bunčić, Sandra L. Lippert, and Achim Rabus, 256–276. Heidelberg: Winter.

Lincke, Eliese-Sophia. 2011. *Die Prinzipien der Klassifizierung im Altägyptischen*. Wiesbaden: Harrassowitz.

Lincke, Eliese-Sophia, and Frank Kammerzell. 2012. "Egyptian Classifiers at the Interface of Lexical Semantics and Pragmatics." In *Lexical Semantics in Ancient Egyptian*, ed. by Eitan Grossman, Stéphane Polis, and Jean Winand, 55–112. Hamburg: Widmaier.

Lincke, Eliese-Sophia, and Silvia Kutscher. 2012. "Motivated Sign Formation in Hieroglyphic Egyptian and German Sign Language (DGS): Towards a Typology of Iconic Signs in Visual Linguistic Systems." In *Lexical Semantics in Ancient Egyptian*, ed. by Eitan Grossman, Stéphane Polis, and Jean Winand, 113–140. Hamburg: Widmaier.

Macramallah, Rizkallah. 1935. *Le Mastaba d'Idout*. Caire: Institut Français d'Archéologie Orientale.

McNeill, David. 2000. "Introduction." In *Language and Gesture*, ed. by David McNeill, 1–10. Cambridge: University Press.

Moussa, Ahmed M., and Hartwig Altenmüller. 1977. *Das Grab des Nianchchnum und Chnumhotep* (Archäologische Veröffentlichungen 21). Mainz: von Zabern.

Müller, Cornelia. 1998. *Redebegleitende Gesten: Kulturgeschichte — Theorie — Sprachvergleich*. Berlin: Berlin Verlag A. Spitz.

Müller, Cornelia, Alan Cienki, Ellen Fricke, Silvia H. Ladewig, David McNeill, and Jana Bressem (eds). 2014. *Body — Language — Communication*. Berlin: de Gruyter.

Norris, Sigrid (ed). 2015. *Multimodality. Critical Concepts in Linguistics*. London: Routledge.

Payrato, Lluís. 2014. "Emblems or Quotable Gestures: Structures, Categories, and Functions." In *Body — Language — Communication*, HSK 38.2, ed. by Cornelia Müller, Alan Cienki, Ellen Fricke, Silvia H. Ladewig, David McNeill, and Jana Bressem, 1474–1481. Berlin: de Gruyter.

Poggi, Isabella. 2014. "Semantics and Pragmatics of Symbolic Gestures" In *Body — Language — Communication*, ed. by Cornelia Müller, Alan Cienki, Ellen Fricke, Silvia H. Ladewig, David McNeill, and Jana Bressem, 1481–1496. Berlin: de Gruyter.

Polis, Stéphane. 2018. "The Functions and Toposyntax of AE Hieroglyphs." *Signata* 9: 291–363.

Polis, Stéphane. 2023. "How Does the Hieroglyphic System Function?" In *Guide to the Writing Systems of Ancient Egypt*, ed. by Stéphane Polis, 42–49. Cairo: Institut Français d'Archéologie Orientale.

Polis, Stéphane, and Serge Rosmorduc. 2015. "The Hieroglyphic Sign Functions." In *Fuzzy Boundaries: Festschrift für Antonio Loprieno*, 2 vols., ed. by Hans Amstutz, Andreas Dorn, Matthias Müller, Miriam Ronsdorf, and Sami Uljas, 149–174. Hamburg: Widmaier.

Raible, Wolfgang. 1978. "Langer Rede Dunkler Sinn. Zur Verständlichkeit von Texten aus der Sicht der Sprachwissenschaft." In *Grammatik und Deutschunterricht. Jahrbuch 1977 des Instituts für deutsche Sprache*, ed. by Ulrich Engel, and Siegfried Grosse, 316–337. Düsseldorf: Pädagogischer Verlag Schwann.

Raible, Wolfgang. 1991. *Semiotik der Textgestalt: Erscheinungsformen und Folgen eines Kulturellen Evolutionsprozesses*. Heidelberg: Winter.

Regulski, Ilona. 2016. "The Origins and Early Development of Writing in Egypt." *Oxford Handbooks Online*. (20/6/2020).

Rogner, Frederik. 2022. *Bildliche Narrativität — Erzählen mit Bildern: Rezeption und Verwendung Figürlicher Darstellungen im Ägyptischen Neuen Reich*. Basel: Librum.

Roth, Ann Macy. 2006. "Little Women: Gender and Hierarchic Proportion in Old Kingdom Mastaba Chapels." In *The Old Kingdom Art and Archaeology: Proceedings of the Conference*, ed. by Miroslav Bárta, 281–296. Prague: Czech Institute of Archaeology.

Royce, Terry D. 1998. "Synergy on the Page: Exploring Intersemiotic Complementarity in Page-Based Multimodal Text." *JASFL Occasional Papers* 1: 25–49.

Royce, Terry D. 2007. "Intersemiotic Complementarity: A Framework for Multimodal Discourse Analysis." In *New Directions in the Analysis of Multimodal Discourse*, ed. by Terry D. Royce, and Wendy L. Bowcher, 63–109. London: Routledge.

Schenkel, Wolfgang. 1984. "Schrift." In *Lexikon der Ägyptologie V*, ed. by Wolfgang Helck, and Wolfhart Westendorf, 713–735. Wiesbaden: Harrassowitz.

Schenkel, Wolfgang. 2003. *Die Hieroglyphische Schriftlehre und die Realität der Hieroglyphischen Graphien*. Sitzungsberichte der Sächsischen Akademie der Wissenschaften zu Leipzig, Philologisch-Historische Klasse 138, 5. Stuttgart/Leipzig/Hirzel: Sächsischen Akademie der Wissenschaften zu Leipzig.

Schenkel, Wolfgang. 2011. "Wie Ikonisch ist die Altägyptische Schrift?" *Lingua Aegyptiae* 19: 125–153.

Seidlmayer, Stephan J. 2012. "Ägyptische Hieroglyphen Zwischen Schrift und Bild." In *Schriftbildlichkeit. Wahrnehmbarkeit, Materialität und Operativität von Notationen*, ed. by Sybille Krämer, Eva Cancik-Kirschbaum, and Rainer Totzke, 123–138. Berlin: Mouton de Gruyter.

Streeck, Jürgen. 2015. "Gesture." In *The Routledge Handbook of Linguistic Anthropology*, ed. by Nancy Bonvillain, 27–43. New York/London: Routledge.

von Stutterheim, Christiane, and Wolfgang Klein. 1989. "Referential Movement in Descriptive and Narrative Discourse." In *Language Processing in Social Context*, ed. by Rainer Dietrich, and Carl F. Graumann, 39–76. Amsterdam: North Holland.

Teßendorf, Sedinha. 2013. "Emblems, Quotable Gestures, or Conventionalized Body Movements." In *Body — Language — Communication*, ed. by Cornelia Müller, Alan Cienki, Ellen Fricke, Silvia H. Ladewig, David McNeill, and Jana Bressem, 82–100. Berlin: de Gruyter.

Thesaurus Linguae Aegyptiae, Korpus-Ausgabe 20, Web-App-Version 2.3.2, 31.10.2025, hrsg. von Tonio Sebastian Richter & Daniel A. Werning im Auftrag der Berlin-Brandenburgischen Akademie der Wissenschaften und Hans-Werner Fischer-Elfert & Peter Dils im Auftrag der Sächsischen Akademie der Wissenschaften zu Leipzig (11/06/2025).

Vandier, Jacques. 1969. *Manuel d'Archéologie Egyptienne. Tome V: Bas-Relief et Peintures. Scenes de la Vie Quotidienne*. Paris: Edition A. et J. Picard.

Vasiljević, Vera. 1998. "Über die Relative Größe der Darstellungen des Grabherrn im Alten Reich." *Studien zur Altägyptischen Kultur* 25: 341–351.

Wertheimer, Max. 1923. "Untersuchungen zur Lehre von der Gestalt, II." *Psychologische Forschung* 4: 301–350.

Wildfeuer, Janina, and John A. Bateman. 2018. "Theoretische und Methodologische Perspektiven des Multimodalitätskonzepts aus Linguistischer Sicht." *IMAGE* 28: 5–46.

Wildfeuer, Janina, Jana Pflaeging, John A. Bateman, Ognyan Seizov, and Chiao-I. Tseng. 2020. "Multimodality: Disciplinary Thoughts and the Challenge of Diversity — Introduction." In *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*, ed. by Janina Wildfeuer, Jana Pflaeging, John A. Bateman, Ognyan Seizov, and Chiao-I. Tseng, 3–38. Berlin/Boston: de Gruyter.

CHAPTER 11

Reflections on literacy in language documentation

Melanie Schippling
University of Cologne

This chapter is concerned with improving our understanding of literacies through language documentation. It addresses different types of writing practices and the use of documentation corpora in this regard. While writing is often taken for granted for languages with a tradition in this modality, it can be seen as "intrusive and alien" (Finnegan 2008: 22) in other contexts. Yet Himmelmann (cf. 2018: 36) points out that hardly any language community nowadays has not encountered a written standard. Accordingly, this chapter argues to increase the efforts in documenting the rich perspectives on and practices of literacy in linguistic communities worldwide. The rationale is that this may enhance knowledge about literate structures in the languages documented as well as in linguistic theory.

Keywords: spoken/written language, writing practices, literacies, language documentation, vernacular writing

1. Introduction

Based on an extensive body of research on spoken and written language, the many ways in which speaking and writing differ, coincide, and relate to each other are acknowledged in linguistics and beyond. Nevertheless, reflections on literacy often remain a side topic in language documentation.[1] Moreover, as Finnegan (2008: 22) puts it, "[…] the established presupposition […] seemed to be that whereas written forms were pretty basic in 'developed' languages, elsewhere writing was somewhat intrusive and alien." However, Himmelmann (cf. 2018: 36) points out that hardly any language community nowadays has not encountered a

1. A note on terminology: I describe the actual practice of producing a text as writing, while the term literacy refers to the broader concept that encompasses writing, reading as well as understanding and creating graphic communicative signs within sociocultural contexts.

https://doi.org/10.1075/slcs.240.11sch
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

written standard. In this line of thought and if writing is used in some form in a community, Finnegan (cf. 2008: 22) argues to take writing into account as authentic linguistic data. Consequently, this chapter addresses the use of language documentation corpora with regard to literacy. The considerations offered also relate to pleas for 'thick description' (e.g. Foley 2003 in analogy to Geertz 1973) and orality and literacy as two out of four levels of sustainable language use against the background of linguistic human rights (e.g. Lewis 2023).[2]

The chapter is structured as follows: I first address the differentiation of orality and literacy and different ways of knowing associated with these modalities. Against this background, Section 2 expounds on the rationale of a comprehensive documentation of writing practices. Section 3 discerns different types of writing practices with regard to a documentation endeavour and, in terms of practical implications for language documentation, further details on examples and proposals for documenting literacy thus far.

2. Orality, literacy, and different ways of knowing

The interrelations of speaking and writing have inspired a considerable body of research. Coulmas (2013:ix) underpins that "[i]f we want to understand the social functions of language, both speech and writing must be taken into consideration, as well as the multifaceted interplay of the two." This thought has proven productive in various ways and can be transferred to the domain of language documentation.[3] In fact, Finnegan (2008: 22) remarks that "insofar as we do recognise close speech-writing ties, then data about this particular form of material representation is indeed relevant in documenting the language".

In the literature, most often, spoken and written language are understood along a continuum (e.g. Koch & Oesterreicher 1985; Maas 2010). Such continua facilitate differentiation of functional characteristics as well as of modalities. For instance, the model of Koch and Oesterreicher (cf. 1985: 23) lists spontaneity and planning among the parameters to discern communicative events along the continuum of communicative distance and proximity. At the same time, it allows to locate a given communicative event regarding its medium of realisation (phonic/graphic). Abstracting away from specific texts, Maas (cf. 2010: 118–119) defines orate and literate structures as not being versus being articulated in the form of sentences respectively and serving a calibrating function for a specific conver-

2. The other levels in this taxonomy proposal are history and identity.

3. For instance, Voßkamp (2010) highlights the oral skills necessary in local journalism to produce high-quality writing.

sation (orate) versus a representation function (literate). It follows from these models that highly structured and elaborate features of a text can be framed as conceptual literacy while literate structures are not restricted to the written modality (consider, e.g., a lecture in a formal context; ritual discourse in an oral tradition). While the individual characteristics of orate and literate structures may differ in linguistic communities worldwide, a different organisation of content depending on modality has implications for language documentation. As Himmelmann (2006a: 272) puts it, "it is important to keep in mind that units in spoken language are often quite different from those in written language."

Regarding the content of language documentation, Seifart (2008: 63) highlights that there is a consensus in documentary linguistics to concentrate on "the use of language in its culture-specific context". I once had a conversation with Nikolaus P. Himmelmann about different ways of knowing that are entailed by different conceptions of literacy. He suggested to compare a 'book-based' literacy with literacy emerging in communities without a long history of writing. For instance, he pointed out that a vast majority of people in Indonesia can read, but the actual gain and exchange of information takes place in the spoken modality (Nikolaus P. Himmelmann, pers. comm., 2021). Similar accounts exist for other contexts such as the following description of a 'written oral literature' in Romani,[4] an Indo-Aryan language with a strong oral tradition: "the overwhelming majority of these texts is intended for oral transmission by recitation or reading. The written form is just a secondary support […]" (Cortiade 1985: 6).

Of course, vernacular writing practices are not necessarily genuinely endogenic. Abdelhay et al. (2014) point out that often, in Africa, they are further developments of and reactions to scripts and practices imported by colonialism. In such cases, a reaction to these imports can develop independently which can also be very interesting to document. Furthermore, while knowledge production may differ even where literacy is understood differently as in 'book-based' cultures, this still does not mean that it should be viewed as "intrusive and alien", as Finnegan (2008: 22) cautions. Literacy may, in fact, be perceived to threaten an oral tradition as Levinson (2007) reports for the case of Romani communities in the United Kingdom. However, Levinson (2007) also finds appreciation of literacy for functional uses on a group level, in a society facing a literacy-based surrounding culture.[5] Against this background, it has proven very productive for language doc-

4. Referring to the Romani language is not meant to ignore that it exists in several dialects without a globally accepted norm. Instead, "[i]t is by far the most widespread term for the language in modern linguistics, and so the most practical cover-term for its various dialects" (Matras 2002: 1). Thus, I use the term to either refer to the entirety of its dialects or to aspects for which a particular dialect is not relevant.

umentation to conceive of literacy as a social practice (Lüpke 2011, see also Sebba 2009).[6] Within the framework of New Literacy Studies, literacies are understood as "multiple, patterned by power relationships, purposeful, historically situated, and changing over time" (Barton, Hamilton, & Ivanič 2000: 2).

Before expounding on the rationale of a comprehensive documentation of writing practices in Section 2, a delimitation should be formulated: This paper is concerned with improving an understanding of literacies through documentation but the point is not to discuss what happens if writing is newly introduced in a community. Several scholars have already pointed out that this may have severe effects (e.g. Ameka 2015, Ong 2002, Dąbrowska 2020). Possible effects pertain to a specific (Western) understanding of literacy and to the way how literacy practices per se change knowledge production. These aspects are briefly outlined in the remainder of this section as they are nevertheless relevant for reflecting on the documentation of different types of writing practices.

Language documentation aims at "a lasting, multipurpose record of a language" (Himmelmann 2006b: 1, emphasis omitted). As such, it can provide resources "serving as a means of establishing and implementing standardised spelling and writing conventions where those are deemed useful by the community" (Lewis 2023: 618). The important point here is the evaluation from the perspective of the community. Regarding a specific understanding of literacy, Western literacy practices indeed favour standardisation and devalue variation. Consider, for instance, how Maitz and Elspaß (2011) point out linguistic discrimination against non-standard varieties in Germany. In such surroundings, there exists a normative understanding of 'good' writing and 'correct' language — a setting Milroy (2001) frames as 'standard language culture'. Thus, "standardisation may be best viewed as an ideology" (Walsh 2021: 774). This ideology may also have consequences if brought unquestioned to a field situation. There are numerous examples of linguists creating orthographies for and with language communities (e.g. the case studies in Jones & Mooney 2020) and it goes without saying that such activities have an impact. Earlier, scholars such as Seifart (2006), Sebba (2009) and Lüpke (2011) expounded on extralinguistic factors of orthography development. Lüpke (2011: 312–313), for instance, highlights that "[…] the potential role and scope of literacy […] in an endangered language needs to be evaluated prior

5. On a socio-political level, different types of culturally accepted forms of knowledge production may cause an imbalance of power. See Witt (2021) for a critique of excluding non-literate knowledge from an academic canon.

6. In this line of thought, understanding literacy as the norm-based technical skill predominantly valued in a 'book-based' culture can be regarded as the manifestation of a specific social practice.

to orthography development." Two decades ago, Seifart (2006: 275) noted that "orthography development is usually not given much attention by linguists." More recently, scholars such as Lüpke (cf. 2017: 142–144) problematise the introduction of a standard into spheres that hitherto function without. Ameka (2015: 27) pointedly states that "standardisation impedes the growth of creative vernacular literacy". Hauk (2022: 30) likewise argues to question an "assumption that a linguist's chief consideration is always the linguistic 'accuracy' of a proposed orthography, while native speakers' motivations are ideological and, by nature, dismissive of scientific validity." In fact, linguists increasingly refrain from imposing the idea of a standard and it is vital and widely acknowledged that the linguistic community should decide how to represent their language in writing.

On a second level, the adoption of literacy entails changes in knowledge production. Ameka (2015: 25) cautions that "the introduction of literacy in a predominantly oral culture imposes new ways of understanding and knowing about the world. […] Over time, literacy brings about a mistrust of the spoken word as evidence." Lüpke (cf. 2011: 321) points out that, with technological progress in audio/video recording, written entities are no longer the only option to share outcomes of a documentation with the community and thus, orthography development and writing in a hitherto unwritten language is no longer necessary unless the explicit need is expressed by the community. Of course, a wish to write in the language for reasons of prestige and emblematic functions is not meant to be ignored by these considerations (cf. also Lüpke 2011).

3. The rationale of a comprehensive documentation of writing practices

Whether against the background of human linguistic rights, with regard to representation, or as a plea for 'thick description' (in analogy to Geertz 1973), several publications address the need to take a comprehensive approach in language documentation (e.g. Himmelmann 1998, 2006b; Foley 2003; Seifart 2008; Lewis and Simons 2016). In this section, I review the main arguments of aiming for comprehensiveness from these perspectives.

Socio-cultural meanings and implications of literacy already received attention in language documentation (e.g. Grenoble & Whaley 1998; Lüpke 2011; Ameka 2015). Moreover, the apparatus and basic/extended format of documentation (Himmelmann 1998, 2006b) do not exclude written entities. Himmelmann (cf. 1998: 180) also discusses a spontaneity parameter in speech to raise awareness for different linguistic practices and their degree of planning. This shows that research on spoken and written language which highlights conceptual differences within the modalities impacted the understanding of language documentation from the start.

However, given the intention to establish documentation within linguistics beyond the then common format of elicited structures and narratives, the focus in these considerations is on audio-visual recordings of observable linguistic behaviour (cf. e.g. Himmelmann 1998: 168; Himmelmann 2006b: 21), i.e. spoken practices. Subsequent criteria and sampling strategies to ensure representativeness (Seifart 2008) also explicitly or implicitly concentrate on spoken language only.

The notion of 'thick description' refers to a dense and grounded description taking structures and backgrounds into account. As such, it aims to understand culture through cultural codes such as rites, practices, or gestures. Foley (2003: 95) points out that such an endeavour necessitates a "very extensive data base". He further adds that "[…] to the extent that our corpus is limited, we straitjacket the knowledge of the language we pass on to future generations, no small failing for the many isolated, politically weak and endangered language communities our work is commonly undertaken with" (Foley 2003: 95).

From another point of view, Lewis and Simons (2016) frame orality and literacy as two out of four levels of sustainable language use for which Lewis (2023) provides an overview of documentation strategies and stakeholders. Lewis (cf. 2023: 617–618) points out that for the other levels, less documentation is needed than for sustainable orality and literacy.

Thus, to be able to provide a 'thick description' aiming at sustainable use of either modality, a solid amount of data is necessary. For spoken language, of course, this has been emphasised in many previous articles in the field of documentation and revitalisation. But regarding sustainable literacy, "[s]imilarly, the corpus must have enough depth to accommodate grammar and discourse studies […]" (Lewis 2023: 618) and different requirements in its documentation may follow. Essentially, the documentation of spoken practices, if there are no written practices that exist prior to the documentation, can lay the foundation for both sustainable orality and literacy. But for achieving sustainable literacy, writing as practised in a community prior to a documentation should play the most important role if it exists. The following paragraphs elaborate on these aspects.

In practice, documentation corpora largely represent spoken language as raw data while theorising and deriving primary data in research often relies on written practices.[7] For instance, "transcription aims to derive primary data (standardized symbolic representations) from raw data (observed linguistic behavior)"

7. The terms raw and primary data are understood in the sense of Himmelmann (2012). Regarding their processing stage, raw data are particular and unique while primary data are particular and non-unique (cf. Himmelmann 2012: 192). Examples (cf. Himmelmann 2012: 193) based on observable linguistic behaviour are recordings or non-standard writing (both raw data) of which primary data are derived by transcription or standardisation.

(Himmelmann 2012: 194). A transcription, however, may introduce a barrier for a community that aims for sustainable orality. Thus, not all approaches rely on spoken interaction that leads to a written documentation and/or analysis. The 'Basic Oral Language Documentation' (BOLD, see Bird 2010) method emphasises the agency of the community. As Himmelmann (2018) points out, transcription is a complex process that involves not just converting speech to text but also understanding the implications of this conversion on language ecology. The BOLD method addresses some of these challenges by focusing on audio(visual) recording by speakers and subsequent speaker-generated oral annotation, including careful respeaking and translation (for details, see Bird 2010). It bridges the gap between orality and literacy not only in making oral literature accessible for both research and community purposes but also in being a method at the intersection of the modalities: Functioning in the spoken modality, it represents conceptual literacy as well. The corpus gained on behalf of this method or selected parts of it can but does not necessarily have to be analysed in the written modality. As such, this approach best caters the needs of a documentation that aims for sustainable orality in the sense of Lewis (2023). Regarding the goal of sustainable literacy, several points need to be considered. First, while writing a language is deemed important for its *Ausbau* ('extension') in the sense of Maas (2010), a "widespread but problematic belief is the necessity for a language to be written in order to be a fully fledged language" (Lüpke 2011: 320). The practice of language documentation to concentrate on spoken language has contributed to establish the status of fully fledged languages for several hitherto unwritten languages. Aiming for sustainable literacy in a documentation context is not meant to ignore this effort and understanding of languagehood. It rather extends the potential means of a language (see also Maas 2010; Lüpke 2011). Second, scholars such as Lüpke (2011) discern endographia from exographia and point out that there are several possible scenarios of writing.[8] If the language of writing is the vernacular spoken by the community, endographia can be documented. If it is different from this vernacular, exographia can further be distinguished into two subtypes:

> 1. a situation where a written majority language with close cultural and/or linguistic affiliations is already present in the multilingual repertoire;
> 2. exographic practices using an official (often ex-colonial) language that is not part of the everyday repertoire of the endangered language community [...].
>
> (Lüpke 2011: 320)

8. These aspects also touch upon the notion of defining an L1 in often multilingual contexts. See Lüpke (2011) for further reflections and Lüpke (2017) for considerations in relation to different models of mono-/multilingualism.

While exographia of the second type can be common and beneficial for the community in terms of practical or socioeconomical considerations, cognitive and psychological benefits have been highlighted for endographia (cf. Lüpke 2011: 319 and the references cited there). As a consequence, the latter may be favoured in arguments against the background of linguistic human rights as it is also deemed to ensure cultural continuity and community empowerment (see, e.g., Akinwalere 2023).[9] Still, as Lüpke (2011: 318) points out, "[e]xographic writing traditions are often overlooked or marginalized […], but it is always worthwhile to conduct a detailed study on functions and uses of writing in other languages prior to embarking on orthography development for an endangered language." In fact, it is increasingly questioned whether orthography development should form an essential part of language documentation (e.g. Lüpke 2011). On the contrary, documenting writing as practised by the community can mean sustainable literacy of the kind that ensures that these practices are not simply *overwritten* by standard language culture. In other words, it is in the interest of both the communities and linguistic theory to document writing practices to allow for a remodelling of literacy in the sense of Lüpke (2017) and achieve a goal of sustainable literacy:

> On a polyglossic plane dominated by Western and monolingual ideas of writing, participation is the main concern in order to achieve visibility, even if successful communication cannot be achieved by these means. […] literacy models need to be rethought if they intend to go beyond the symbolic. (Lüpke 2017: 141)

Beyond documenting genres and describing how language is represented in writing, a thick description of literacy in any community should also include data on the use of written artefacts, the way they are distributed, and their reception. A typology of written artefacts in a language community may also be very informative about functional domains of language use.[10] Language documentation often focusses on code-switching in the spoken domain (e.g. Reetz 2022) and is rightly

9. Note that while learning to read in an L1 generally supports cognitive and literacy development, the benefits may vary based on the context and implementation. Some studies show no significant advantage in second language learning outcomes (e.g. Van Wyk & Mostert 2016; Piper et al. 2018), while others highlight a positive transfer of skills in contexts of both the Global North and South (e.g. Taylor & von Fintel 2016; Leroy et al. 2021; Gutiérrez-Fresneda et al. 2024).

10. For instance, there are stitched and unstitched Islamic manuscripts in West Africa whereby the latter discontinues the original tradition (Bondarev 2023). Formulating hypotheses on the emergence of this differentiation, Bondarev (2023: 296) suggests that "the unsewn type of binding was restricted to a specific functional domain: learning and memorising the Qur'an." It remains to be reconstructed how this narrowly defined functional domain was expanded so that it features in all stages of Islamic education now.

interested in underlying mechanisms and language contact. In fact, the written domain may be as insightful. For instance, Bondarev (2021: 718) describes a use of both Arabic and a vernacular in Ajami[11] manuscripts: "When knowledge of Arabic is insufficient by the scribe, he switches to familiar terms in the language he knows best, the process resembling codeswitching in speech." Bondarev (2021) proposes a typology for West African Ajami manuscripts which allows to systematically investigate the functional domains of the languages used in different types of manuscripts.

It follows from these considerations that promising insights can be gained from comprehensive documentation of writing practices. Even if written language appears to be more durable than spoken language, the issue is partly as urgent: One premise of the global documentation efforts is the time pressure in view of the decreasing diversity of languages worldwide. Linguists are rushing to record the language use of last speakers. Where writing and reading practices change, such a time pressure may equally exist. An example are traditional Tai Ahom manuscripts which an older generation is still able to read whereas younger generations are not trained to do so (cf. Morey 2015: 52). With regard to feasibility, Morey (cf. 2015: 52–53) describes the time-consuming creation of even a basic description of one manuscript in terms of metadata which requires in-depth knowledge and experience.

In fact, the many options of focus and feasibility pose further challenges. Being aware that not every instance of language use can be documented, Seifart (2008: 60) raises the question of how to achieve representativeness: "Imagine that at some later stage of history, the data available about our western-style civilization would be heavily biased towards (or even limited to) Grimm's fairy-tales (or soap operas, or theoretical physics)." He proposes to discern three sampling strategies to address this issue: 'convenience sampling', 'externally motivated sampling', and 'systematic sampling' (Seifart 2008: 64). These strategies add to the apparatus and basic/extended format of documentation proposed by Himmelmann (1998; 2006b) and concentrate on the content of the communicative events (in the sense of Himmelmann 1998: 168) to be documented (cf. Seifart 2008: 64). The systematic sampling approach then allows to choose criteria or parameters based on a framework in order "ensure the representativeness of a documentation in a theoretically grounded way" (Seifart 2008: 60). It involves categorising event types according to, e.g., the spontaneity parameter (in the sense of Himmelmann 1998) along a continuum (cf. Seifart 2008: 65–66). This continuum of 'planned' – 'unplanned' (cf. Seifart 2008: 66) is reminiscent of continua conceptualising spoken and writ-

11. The term Ajami refers to the use of the Arabic script for non-Arabic languages, particularly for writing African languages (Lüpke 2017; Bondarev 2021).

ten language — yet the latter open up a multi-dimensional space. (Weighted) frequency criteria (cf. Himmelmann 1998: 181, Seifart 2008: 69) and further sampling approaches, such as methods from the ethnography of communication (see Seifart 2008: 66–67) of course also allow for an integration of writing practices into a documentation and fit "well in the broad perspective of language use in its cultural context that is the object of language documentation" (Seifart 2008: 66). Notwithstanding the awareness of conceptual orality or literacy in the spoken modality present from the beginning of theorising language documentation, it seems worthwhile to (again) additionally consider models of spoken and written language as a framework in a systematic sampling in order to integrate the additional modality throughout. Seifart (cf. 2008: 69) acknowledges possible biases in analytical decisions setting up a systematic sampling approach and Finnegan (2008: 26) cautions that "[l]anguage documentation can never be, has never been, a matter of detached and objective pebble gathering, but an intensely human process of selection, analysis and, inevitably, manipulation." Now might be a good time to further develop the documentation of writing practices empirically and theoretically. The rationale is that the data, in turn, may allow for an adaption of precisely the models of orality and literacy existing thus far and contribute to an improved understanding of literacies worldwide (as advocated by Lüpke 2017).

4. Different types of writing practices, examples and implications for their documentation

As noted in Section 2, language documentation most often concentrates on spoken language. Writing practices may be included but often represent a side note. For instance, in describing the approach of documenting a language, Austin (cf. 2014: 60) only includes written samples in parentheses. This section highlights some examples, proposals and implications for documenting writing practices.

A question related to sampling criteria in this context is whether certain traces of language in the graphic domain count as writing. This pertains especially to historical linguistics and investigations of prehistoric manuscripts. But there are, for instance, also examples of non-alphabetic signs used in Romani communication in the nineteenth century and more recent times (cf., e.g., Toninato 2014: 57–58). Bausi et al. (2023: 1) define a 'written artefact' as "any artificial or natural object with visual signs applied by humans." This broad definition covers various writing practices over the course of history and the authors suggest to also reflect on materiality, possibilities for interaction and transmission of a given written artefact (cf. Bausi et al. 2023: 7). Such considerations tie with the notion of knowledge production as elaborated in Section 1 and can be used as a starting point in a documentation endeavour. "[F]rom absolute ephemerality (writing on

sand) to preservation forever; from temporary use, as in the case of a school book, to the archival preservation for occasional consultation" (Bausi et al. 2023:7), there are many varieties of performing or producing knowledge.

4.1 Documenting existing writing practices

Written practices existing prior to any documentation endeavour without a doubt count as "linguistic practices characteristic of a given speech community" Himmelmann (1998:166) to be documented. In fact, writing and literacies have been analysed for diasporas or communities beyond contexts of an established written tradition. However, such studies often take an ethnographic (e.g. Abercrombie 2018; Mbodj-Poye 2013) rather than linguistic approach. Yet linguistic research can strongly benefit from including information on literacy in documentation efforts. Consider the remarks of Lüpke (2017:132–133) on the case of West Africa:

> [...] the lack of coverage of rural multilingual settings, which have been shaped under very different social conditions, and in which the official languages are marginal, also means that the development of models for literacy in national languages of Africa did not have the chance to be inspired by these settings. The development of models adapted to the full scale of multilingualism must therefore be of prime importance [...].

Thus, documenting writing practices may indeed pave the way for a renewed understanding of literacy given that a lack of descriptions and case studies often leads to normative understandings of literacy filling the void.[12] Lüpke (2017:129) frames writing as practised by a community beyond a standard language culture as a "culturally anchored way of writing". In fact, this may co-exist with a codified form of literacy which may be adapted to the communicative needs of the community, a concept Lüpke (2017:134) circumscribes as lead language writing:

> Just as Ajami writing is seldom a standalone literacy, grassroots literacies in the Latin script have a lead too. One and the same language can be written with very different sound-grapheme associations depending on which language provides the orthographic conventions serving as the lead.[13]

---

12. For a transfer of literate structures known from a dominant language to writing in another language even in the context of two languages with a written tradition, see Schroeder (2014).

13. For an overview of Ajami as a *non-governmental literacy*, see Souag (2010). Its numerous adaptions and widespread use can make it difficult to identify the language of a manuscript. Bondarev et al. (2019) provide fixed correspondences and diagnostic features to alleviate this task. Based on an increased documentation of writing practices, such resources may become available for further scenarios.

The notions of a *linguistic repertoire* (Busch 2012) or a *language as resource* matrix (Coulmas 2013) represent further proposals to encompass a holistic view of both an individual's and a community's resources and patterns of language use. These concepts are united by their emphasis on fluidity and adjustment. For instance, Lüpke (2017: 133) highlights the performativity aspect of "adaptive conventionalization" and argues that "[t]he potential for signs to express social meaning is dependent on making variation meaningful, and hence on regimes that tolerate it and do not insist on a standard."

Documenting writing practices that exist prior to a documentation may of course also cover language use on social media. So-called grassroots literacy flourishes on such platforms and social media are now seen as an asset for language revitalisation (cf. e.g. Jany 2017). But while using endangered languages on social platforms is advocated and the role of language use on these platforms is analysed by several language preservation endeavours (e.g. Riyanto et al. 2023; Lidubwi and Ndavula 2024; Minhas and Salawu 2024), a documentation of actual writing practices on social media beforehand is less frequently found: Tobar (2019: 143) states that this data type is "still uncommon in the field of language documentation and description". A rare example is Handoko (2024) who aims at diversifying the available corpus of Minangkabau, an Austronesian language of Indonesia, by including data from social media. In this way, informal online communication may also become available for analysis. In other words, writing on social media may be incorporated into a documentation as raw data (in the sense of Himmelmann 2012). Tobar (cf. 2019: 143) equally advocates to treat such online language samples as raw data in a case study on a creole spoken in the Philippines and underlines that

> for many of the speakers of this creole language, a great deal, if not all, of their writing practices take place online [...]. Social media networks can provide researchers with an important source of contemporary linguistic data in the case of poorly described languages like Zamboanga Chabacano, as they offer ever-increasing amounts of naturally occurring language. (Tobar 2019: 165)

An online documentation of non-standard varieties may also open up a longitudinal perspective: Analyses show that a "new acceptance of regional features in the standard variety is related to the prevalence of social media" (Zhao & Liu 2021: 884) due to its discursivity and diversity of voices (cf. Zhao & Liu 2021: 884 with references to Heller 2009 & Eckert 2008).

4.2 Documenting writing practices that emerge through documentation

In the context of a documentation, writing takes place in some form or another. Written practices resulting from it (e.g. transcriptions, orthography development) should themselves also be a subject of the documentation as these practices allow an analysis of both metalinguistic knowledge and potential reactions to different concepts of literacy. Moreover, their inclusion contributes to research transparency.

In their workflow, linguists often collaborate with various transcribers from the language community. These transcribers may employ different writing conventions or even offer scripts that showcase inter-individual differences and variability. In contrast, there is often an established way of representing a specific language in academia which will likely be used by linguists writing up an academic report. For instance, there is a writing system commonly used in Romani linguistics as opposed to pluralistic practices employed by Romani individuals (see Schippling 2026) or an orthography established by Parker and Parker (1974) for the non-Austronesian Baining language Qaqet as opposed to different notations used by transcribers (see Hellwig 2019; Reetz 2022:32; see also Walker et al., this volume). Additionally, conventions for spelling separately may differ as in Example (1) from the Qaqet documentation. The notation frequently found in transcripts done by speakers of Qaqet is represented in (1a) with a separate spelling of the constituents. Hellwig opts for the combined spelling in her transcription (1b) as she analyses the Qaqet pronominal arguments as suffixes since they "invariably form a phonological word with the preceding preposition or verb" (Hellwig 2019: 98) and "trigger phonological changes in preceding elements that are otherwise only attested in suffixes, not in enclitics" (Hellwig 2019: 99).

(1) a. *kurli ngia*
b. *kurlinya*[14]
kurli-nyi=a
leave-2SG=DIST
'you stay now' (LONGYDS20150531_1-507/transcript original number 364)

A further difference is the spelling of /ɲ/ as <ng> (1a) and <ny> (1b): Hellwig (cf. 2019: 34) analyses /ɲ/ as contrastive to /ŋ/ whereas the orthography represents both as <ng>. To enable a precise representation of her linguistic analysis, thus, she departs in a transparent manner from the orthography and the word boundaries as found in the transcripts. However, to accommodate the notation familiar to the speakers of Qaqet, she utilises <ng> throughout in a storybook published

14. Example, glossing, and translation adopted from Hellwig (2019: 242).

with and for the community but utilises combined spelling here as well (e.g. *qurlingi*, see Frye & Hellwig 2022).

While a choice has to be made in such cases and the choice itself may be perfectly reasonable, it presents researchers with a dilemma: The conventions chosen strengthen one specific way of writing and may devalue others. Taking these considerations further, Foley (2003: 95) acknowledges that "[o]ur work will be normative for these communities. Unfortunately, what we describe of the threatened languages of many small language communities is often all that will be passed on of them to their descendents. We owe it to them to provide the 'thickest description' we are capable of."

Several linguists have made an effort to produce written texts that align with local practices rather than normative literacy traditions. For instance, in documenting Teop, an Austronesian, Oceanic Western Melanesian language spoken in Papua New Guinea, Mosel's (cf. 2015: 323) efforts of producing written texts for a hitherto unwritten language aligned with the wishes of the speech community. To avoid Westernisation, Mosel (cf. 2015: 323) gave instructions for the editing of transcribed recordings that aim at minor changes, but the editing went beyond these rules. She denominates the texts produced as new text types created in the realm of the documentation which led to the evolution of a Teop text production (cf. Mosel 2015: 323). Mosel (2015: 340) highlights the scientific benefits of the texts produced: These "give a fuller picture of the use of a wide range of syntactic constructions; […] show alternative ways of expressing the same content; […] provide a new type of data for research on the preferred lexical and grammatical features of spoken and written registers in comparable corpora."

That is, documenting such writing practices may enhance knowledge about orate and literate structures of a language and linguistic theory at the same time. However, as noted in Section 2, papers concerned with representativeness of documentation efforts concentrate on spoken language only (e.g. Seifart 2008; Muradoglu et al. 2023). This observation in no way diminishes the value of the careful considerations made in these articles, but merely shows that another type of data is not (or rather cannot yet be) considered. Put another way, in leaving out writing practices, an informative part of language use remains undocumented.[15] My argument follows a similar line of thought as those of Hellwig and Jung (2020). With a different focus of expanding documentation endeavours, Hellwig and Jung (2020: 188) state that "child language and child-directed language constitute data types that can inform our understanding of the adult language. […] this data type provides insights into the metalinguistic knowledge of adult

15. As a result, misunderstandings can arise, even in communication intended to be tailored to the target audience (see Lüpke 2017: 139–140).

speakers". Likewise, writing may function as a window into the metalinguistic knowledge of speakers.[16] Documenting vernacular writing thus goes beyond comparing preferred structures across modalities. Moreover, while including vernacular writing in the domains of language maintenance and revitalisation appears more straightforward, it is language documentation which could inform the most authentic baseline. Documenting emerging writing practices would also enable a retrospective analysis of the extent to which potential unintended consequences of introducing a certain concept of literacy resulted from the documentation (see Section 1).

4.3 Prospects of documenting writing practices that exist prior to or emerge through documentation

It is widely acknowledged that documentary linguistics is not restricted to collecting data but instead "is in a position to make substantive contributions to linguistic theory" (Good 2018: 13, see also Himmelmann 2012; Hellwig & Jung 2020). This equally pertains to documenting writing — both existing prior to or emerging in the context of a documentation — as raw and primary data: As elucidated in Sections 3.1 and 3.2, documenting both allows insights into orate and literate structures of a diverse range of languages. Beyond scientific interest, this may also have (possibly positive) consequences for diversity in further spheres.

In writing up a research report, linguists have to choose a way to represent the language analysed in writing which often involves deciding which orthography is the best fit for academic discourse about it. We can look at this decision as we look at the selection of relevant data for analyses. In this context, Finnegan (2008: 26, o.e.) writes: "Selective choices are inevitable. But we should be clear that we *are* making them, that by going down one route we are excluding others, and — finally — that our decisions about what data counts may mean in effect tacitly lining up with some particular position about the nature and working of 'language'." In other words, choosing a single way of representing a language reduces variation and tacitly entails endorsing the idea of a standard. One could argue that if the variation encountered in a documentation were incorporated into academic papers, comparability within the research field would no longer be possible. To address this, it might be a compromise to work with different tiers in providing examples. Using different tiers is perfectly normal in transcription software in documentary linguistics and this working mode can be transferred to descriptive linguistics.

---

16. See also Himmelmann (2006a: 257) who discusses metalinguistic knowledge conveyed in a separate or combined spelling of compounds.

To contextualise this view, consider Example (2) from Romani. Despite being spoken in a literacy-based surrounding with major groupings in Europe, Australia, South America, there is no standard form of the language in use. From scripts to spelling, the notation of Romani shows a considerable degree of variation. The notation in academic publications, on the other hand, is rather coherent. In this situation, both versions may be utilised in an academic report as in Schippling (2026): Handwritten texts by speakers of Romani were typed exactly as written (2a) while spoken data was transcribed using the orthography commonly found in Romani linguistics (2b). Thus, an influence of the Albanian orthography, another L1 of the speaker, becomes apparent in the representation of /tʃh/ as <çh> (2a) whereas the academic conventions propose a use of <čh> (2b).

(2) a. *E çhaves isine taj jek Rikono*
ART boy.OBL was.3SG CONJ ART dog
b. *e čhaves isine jekh rikono*
ART boy.OBL was.3SG ART dog
'The boy also had a dog' (OEFFROG)[17]

This dual example structure serves three purposes: First, it does not overwrite the variability in the data. Second, it does not give priority for any notation. And third, using the orthography established in academia serves comparison with scholarly publications. Schippling (2026) is admittedly concerned with the interrelations of speaking and writing Romani, however, this approach can easily be adapted to any research interest.

Similar approaches to using multiple tiers in this way are proposed with regard to translation. For instance, Macht (2018:§10) argues to include words of another language in a translated text as this "breaks the dominion of the English language if only temporarily. It re-orients the reader's attention to there being a textual difference that needs to be attentively considered". Likewise, including variation in our representation of language serves to break the dominion of the idea of a standard.

Moreover, while transcription most often involves editing and further decision-taking, documenting writing as practised in the community does not need to be biased by these processes. Schultze-Berndt (2006: 219) pointed out these advantages two decades ago and likewise addresses a similar possibility of a dual mode of written representation in a documentation: "Written communicative events usually employ an orthographic representation (which may or may not be standardized; in the latter case, a rendition in standardized orthography could be added

17. Example from the Mečkara dialect of Romani. See Schippling (2026) for more information on its source based on a Frog story (Mayer 1969) re-narration.

to the documentation)." Unprompted and unbiased written entities may then also serve as valuable points of reference for transcription practices in a documentation. For instance, Klamer and Moro (2020) compare the 'naturalness' of free narratives and narratives prompted by the Frog story (Mayer 1969) in two hitherto unwritten languages along different variables such as lexical density or tail-head linkage. In analogy to this study, a comparison of writing practised in a community independently of a documentation and writing emerging from it might enable to identify similar linguistic variables to determine the naturalness of written communicative events. This would allow for a more thorough reflection on the documented modality and shed light on how written data emerging from a documentation relate to writing as practised by the community.

Even though the degree of influence caused by a normative literate tradition remains hard to determine, the efforts described by Mosel (2015, see Section 3.2) also align with a suggestion by Ameka (2015: 32) to "develop ways of communicating meanings in diverse ways (even in written texts) that preserve variation and reduce the gulf between spoken language and written language by promoting more vernacular literacy." In fact, there is also a 'new culture of vernacular literacy' (Elbow 2012) or, put differently, an emerging trend of 'de-standardisation' in standard language cultures which entails an increasing recognition of variability (cf. e.g. Mattheier 2003: 239–241; Zhao & Liu 2021). Such processes are currently accelerated by language use on the internet in general and social media in particular (Elbow 2012, Zhao & Liu 2021).

However, online communication often remains impacted by standard language ideology (cf. Zhao & Liu 2021: 892).[18] The issues this entails are exacerbated by an increasing use of natural language processing (NLP) techniques in appliances of artificial intelligence (AI). At the present time, scholars discuss to what extent NLP techniques may be of use to address the transcription bottleneck (in the sense of Himmelmann 2018, Seifart et al. 2018) in language documentation (e.g. Spencer 2024). However, just as a majority of research efforts continue to concentrate on the major languages of the world, Large Language Models (LLMs) are predominantly trained on these languages — and against the background

18. As one anonymous reviewer rightly pointed out, this does not necessarily generalise across all settings of language use worldwide. In fact, for the use of two local Putonghua varieties on a microblogging platform analysed by Zhao and Liu (2021), a direct comparison with a standardised form is possible and an influence of a monolingual language policy is not surprising. Yet, even in the case of Romani with no generally accepted form of a standard in use, other standardised forms of language are often seen as a model, especially for written communication (cf. Schippling 2026). Moreover, Abercrombie (2018: 746, o.e.) observes in a study on language purism in a Romani community that "the IDEA of a standard orthography is ever-present." The degree of possible impacts of course varies per context.

of literacy models known from standard language cultures. Hence, in a cross-cultural analysis of the performance of LLMs, Shan et al. (2025: 106) find "biases and inaccuracies in languages with less digital representation. These biases not only affect the accuracy of the models but also potentially exacerbate existing social inequalities […]." Even for an everyday use of English, talks such as Schlüter and Deckert (2025) on AI-assisted writing showcase a growing awareness that AI might overwrite variability in non-standard forms of language. Different initiatives aim to counteract this situation.[19] In including writing practices in language documentation, there is the potential to train neuronal networks on new models of literacy in the future. As Lewis (2023: 616) puts it: "While LD is often backward looking — an attempt to preserve a past that is being lost — it also has very practical forward-looking applications when the community leads."[20]

5. Conclusions

To summarise, linguistic studies of writing as practised by a community are largely lacking within the domain of language documentation. As the edited volume that this paper features in underlines, putting data centre stage is an important merit of language documentation. However, it would be a missed opportunity if this could only be said regarding spoken language. Linguistic theory as well as non-academic stakeholders may equally benefit from more diverse data on writing practices. This concerns several levels of linguistic analyses such as spelling or word boundaries and may be informative of metalinguistic knowledge. Moreover, such an expanded scope of the documentation would allow to render transparent how notation and analyses in linguistic research relate to writing as practised by a language community. While documenting literacies can be an epistemological interest per se, it also contributes to maintain diversity. Archiving such data and reporting case studies are a valuable start which may inform new models of literacy and, sub-

19. For instance, Masakhane (https://www.masakhane.io/home) promotes NLP research in African languages; AmericasNLP (https://turing.iimas.unam.mx/americasnlp/index.html) concentrates on native American languages, and digiserb (https://digiserb.de/de/) is concerned with digital topics around Sorbian languages spoken in Germany.

20. This quote places an important emphasis on the communities' linguistic rights: Using data from language archives for training algorithms, commercial entities must carefully manage ethical and legal questions. To date, research explores the applicability of LLMs on protecting personal data in archives (e.g. Yang et al. 2023). There are also corpora explicitly built for offering training material for NLP techniques in response to an underrepresentation, such as the state-funded speech corpora by the Linguistic Data Consortium for Indian Languages (see Choudhary & Rao 2020).

sequently, tailored teaching materials. Beyond traditional applications of documentation materials, NLP techniques could alleviate transcription processes in the future based on the data. Moreover, AI may be trained on new models of literacy so as not to leave the field to standardisation. That is, in an optimistic view, technological progress, new models of literacy, and representation of small languages in further domains could go hand in hand.

Acknowledgements

For feedback and comments on earlier versions of this paper, I wish to thank Birgit Hellwig, Uta Reinöhl, and an anonymous reviewer. All remaining errors are my own.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following glosses and abbreviations are used:

AI	artificial intelligence	LLM	large language model
CONJ	conjunction	NLP	natural language processing
L1	first language	o.e.	original emphasis.
LD	language documentation		

References

Abdelhay, Ashraf, Yonas Mesfun Asfaha, and Kasper Juffermans. 2014. "African Literacy Ideologies, Scripts and Education." In *African Literacies. Ideologies, Scripts, Education*, ed. by Kasper Juffermans, Yonas Mesfun Asfaha, and Ashraf Abdelhay, 1–62. Newcastle upon Tyne: Cambridge Scholars.

doi Abercrombie, Amelia. 2018. "Language Purism and Social Hierarchies: Making a Romani Standard in Prizren." *Language in Society* 47 (5): 741–761.

doi Akinwalere, Ifedayo. 2023. "Indigenous Language as a Factor for Sustainable Development." In *Indigenous African Language Media*, ed. by Phillip Mpofu, Israel Ayinla Fadipe, and Thulani Tshabangu, 321–337 Singapore: Palgrave Macmillan.

doi Ameka, Felix K. 2015. "Unintended Consequences of Methodological and Practical Responses to Language Endangerment in Africa." In *Language Documentation and Endangerment in Africa*, ed. by James Essegbey, Brent Henderson, and Fiona McLaughlin, 1–15. (Culture and Language Use: Studies in Anthropological Linguistics 17.) Amsterdam: Benjamins.

doi Austin, Peter K. 2014. "Language Documentation in the 21st Century." *JournaLIPP* 3: 57–71.

doi Barton, David, Mary Hamilton, and Roz Ivanič (eds.). 2000. *Situated Literacies. Theorising Reading and Writing in Context*. (Literacies 1.) London: Routledge.

Bausi, Alessandro, Dmitry Bondarev, Nadine Bregler, Sara Chiarini, Giovanni Ciotti, Janine Droese, Eliana Dal Sasso, Michael Friedrich, Agnieszka Helman-Ważny, Michael Kohs, Leah Mascia, Ann Lauren Osthof, Malena Ratzke, Ondřej Škrabal, Szilvia Sövegjártó, Thies Staack, and Hanna Wimmer. 2023. *Definition of 'Written Artefact'*. (CSMC Occasional Paper 8.)

Bird, Steven. 2010. "A Scalable Method for Preserving Oral Literature from Small Languages." In *The Role of Digital Libraries in a Time of Global Change*, ed. by Gobinda Chowdhury, Chris Koo, and Jane Hunter, 5–14. Berlin: Springer.

Bondarev, Dmitry. 2021. "A Typology of West African Ajami Manuscripts: Languages, Layout and Research Perspectives." In *Exploring Written Artefacts: Objects, Methods, and Concepts*, ed. by Jörg B. Quenzer, 707–728. Berlin, Boston: De Gruyter Mouton.

Bondarev, Dmitry. 2023. "Loose-leaf Islamic Manuscripts of West Africa: Retention, Adaptation or Invention?" In *Tied and Bound: A Comparative View on Manuscript Binding* ed. by Alessandro Bausi, and Michael Friedrich, 285–299 Berlin, Boston: De Gruyter Mouton.

Bondarev, Dmitry, Nikolay Dobronravin, Darya Ogorodnikova, and Lameen Souag. 2019. "Identifying Languages of Ajami Commentaries." *Zenodo*.

Busch, Brigitta. 2012. "The Linguistic Repertoire Revisited." *Applied Linguistics* 33 (5): 503–523.

Choudhary, Narayan, D.G. Rao. 2020. "The LDC-IL Speech Corpora." In *Proceedings of the 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques*, Yangon, Myanmar. 28–32.

Coulmas, Florian. 2013. *Writing and Society. An Introduction*. Cambridge: Cambridge University Press.

Courthiade, Marcel. 1985. "Between Oral and Written Textuality: The Lila of the Young Romani Poets in Kosovia." *Lacio Drom* 6: 2–20.

Dąbrowska, Ewa. 2020. "How Writing Changes Language." In *Language Change: The Impact of English as a Lingua Franca*, ed. by Anna Mauranen, and Svetlana Vetchinnikova, 75–94. Cambridge: Cambridge University Press.

Eckert, Penelope. 2008. "Variation and the Indexical Field." *Journal of Sociolinguistics* 12 (4): 453–476. . (Cited after Zhao & Liu. 2021.)

Elbow, Peter. 2012. *Vernacular Eloquence: What Speech Can Bring to Writing*. Oxford: Oxford University Press.

Finnegan, Ruth. 2008. "Data — but Data from What?" In *Language Documentation and Description* 5: 13–28.

Foley, William A. 2003. "Genre, Register and Language Documentation in Literate and Preliterate Communities." *Language Documentation and Description* 1: 85–98.

Frye, Henrike, and Birgit Hellwig (eds.). 2022. Arluis araataing ngenamasiit [*Children's songs and stories*]. Plauen: Sachsendruck.

Geertz, Clifford. 1973. *The Interpretation of Cultures*. New York: Basic Books.

Good, Jeff. 2018. "Reflections on the Scope of Language Documentation." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 13–21. (Language Documentation and Conservation Special Publication 15.) Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24804

Grenoble, Lenore A. & Lindsay J. Whaley. 1998. "Toward a Typology of Language Endangerment." In *Endangered Languages: Language Loss and Community Response*, ed. by Lenore A. Grenoble and Lindsay J. Whaley. Cambridge: Cambridge University Press. 22–54.

Gutiérrez-Fresneda, Raúl, Teresa Pozo Rico, María José García Tárraga, and Elena Jiménez-Pérez. 2024. "Effect of the Development of Precursor Skills for Literacy in Spanish on Learning to Read in English." *The Journal of Educational Research* 117 (5): 284–293.

Handoko, Handoko. 2024. "Developing the Corpus of Minangkabau Language: Insights, Challenges, and Future Directions." *Jurnal Arbitrer* 11 (3): 413–429.

Hauk, Bryn. 2022. "Review of Creating Orthographies for Endangered Languages, ed. by Mari C. Jones, and Damien Mooney." *Language Documentation and Conservation* 16: 25–31. http://hdl.handle.net/10125/74674

Heller, Monica. 2009. "Media, the State and Linguistic Authority." In *Language Ideologies and Media Discourse: Texts, Practices, Politics*, ed. by Sally Johnson, and Tommaso M. Milani, 277–282. London: Bloomsbury Academic. (Cited after Zhao & Liu. 2021.)

Hellwig, Birgit. 2019. *A Grammar of Qaqet*. (Mouton Grammar Library, 79). Berlin, Boston: De Gruyter Mouton.

Hellwig, Birgit, and Dagmar Jung. 2020. "Child-directed Language — and how it Informs the Documentation and Description of the Adult Language." *Language Documentation and Conservation* 14: 188–214. https://hdl.handle.net/10125/24920

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2006a. "The Challenges of Segmenting Spoken Language." In *Essentials of language documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 253–274. Berlin, New York: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2006b. "Language Documentation: What is it and What is it Good for?" In *Essentials of language documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin, New York: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2012. "Linguistic Data Types and the Interface between Language Documentation and Description." *Language Documentation and Conservation* 6: 187–207. http://hdl.handle.net/10125/4503

Himmelmann, Nikolaus P. 2018. "Meeting the Transcription Challenge." In *Reflections on Language Documentation. 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 33–40. (Language Documentation and Conservation Special Publication 15.) Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24806

Jany, Carmen. 2017. "The Role of New Technology and Social Media in Reversing Language Loss." *Speech, Language and Hearing* 21 (2): 73–76.

Jones, Mari C., and Damien Mooney (eds.). 2020. *Creating Orthographies for Endangered Languages*. Cambridge: Cambridge University Press.

Klamer, Marian, and Francesca R. Moro. 2020. "What is 'Natural' Speech? Comparing Free Narratives and Frog Stories in Indonesia." *Language Documentation and Conservation* 14: 238–313. http://hdl.handle.net/10125/24921

Koch, Peter, and Wulf Oesterreicher. 1985. "Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte." *Romanistisches Jahrbuch* 36 (1): 15–43.

Leroy, Lieve Mieke R., Van Phuc Le, Wouter Boesman, and Filip Lenaerts. 2021. "Talking With Children in Multilingual Preschools in Central Vietnam." *Childhood Education* 97 (4): 24–33.

Levinson, Martin P. 2007. "Literacy in English Gypsy Communities: Cultural Capital Manifested as Negative Assets." *American Educational Research Journal* 44 (1): 5–39.

Lewis, M. Paul. 2023. "Promoting Linguistic Human Rights Through Language Documentation." In *The Handbook of Linguistic Human Rights*, ed. by Tove Skutnabb-Kangas, and Robert Phillipson, 613–621. Hoboken: Wiley Blackwell.

Lewis, M. Paul, and Gary F. Simons. 2016. *Sustaining Language Use: Perspectives on Community-Based Language Development*. Dallas: SIL International.

Lidubwi, Jackline U., and John O. Ndavula. 2024. "Revitalising Endangered Languages through Social Media. A Case Study of Olunyore Language Preservation through Facebook in Kenya." In *Decolonising Digital Media and Indigenisation of Participatory Epistemologies*, ed. by Fulufhelo Oscar Makananise, and Shumani Eric Madima, 163–174. London: Routledge.

Lüpke, Friederike. 2011. "Orthography Development." In *The Cambridge Handbook of Endangered Languages*, ed. by Peter K. Austin, and Julia Sallabank, 312–336. (Cambridge Handbooks in Language and Linguistics.) Cambridge: Cambridge University Press.

Lüpke, Friederike. 2017. "Escaping the Tyranny of Writing: West African Regimes of Writing as a Model for Multilingual Literacy." In *Tyranny of Writing: Ideologies of the Written Word*, ed. by Constanze Weth, and Kasper Juffermans, 129–148. London: Bloomsbury Academic.

Maas, Utz. 2010. "Literat und Orat. Grundbegriffe der Analyse geschriebener und Gesprochener Sprache." *Grazer Linguistische Studien* 73: 21–150.

Macht, Alexandra. 2018. "From Romanian 'Soul' to English 'Heart': Dilemmas of Cultural and Gender Representation in Translating Qualitative Data." *Forum Qualitative Sozialforschung* 19 (2): Article 12.

Maitz, Péter, and Stephan Elspaß. 2011. "Zur Sozialen und Sprachpolitischen Verantwortung der Variationslinguistik." In *Dynamik des Dialekts – Wandel und Variation*, ed. by Elvira Glaser, Jürgen E. Schmidt, and Natascha Frey, 221–240. Stuttgart: Steiner.

Matras, Yaron. 2002. *Romani. A Linguistic Introduction*. Cambridge: Cambridge University Press.

Mattheier, Klaus J. 2003. "German." In *Germanic Standardizations: Past to Present*, ed. by Ana Deumert, 211–244. Amsterdam: Benjamins.

Mayer, Mercer. 1969. *Frog, Where Are You?* New York: Dial Books for Young Readers.

Mbodj-Pouye, Aïssatou. 2013. "Writing the Self in Rural Mali: Domestic Archives and Genres of Personal Writing." *Africa: The Journal of the International African Institute* 83: 205–226.

Milroy, James. 2001. "Language Ideologies and the Consequences of Standardization." *Journal of Sociolinguistics* 5 (4): 530–555.

Minhas, Shahid, and Abiodun Salawu. 2024. "Preserving and Promoting Indigenous Languages: Social Media Analysis of Punjabi and Setswana Languages." *Journal of Asian and African Studies* 60 (7): 4109–4126.

Morey, Stephen. 2015. "Metadata and Endangered Archives: Lessons from the Ahom Manuscripts Project." In *From Dust to Digital: Ten Years of the Endangered Archives Programme*, ed. by Maja Kominko, 31–66. Cambridge: Open Book Publishers.

Mosel, Ulrike. 2015. "Putting Oral Narratives into Writing — Experiences from a Language Documentation Project in Bouganville, Papua New Guinea." In *Language Contact and Documentation — Contacto lingüístico y documentación*, ed. by Bernard Comrie, and Lucía Golluscio, 321–342. Berlin, Boston: De Gruyter Mouton.

Muradoglu, Saliha, Hanna Suominen, and Nicholas Evans. 2023. "A Quest for Paradigm Coverage: The Story of Nen." In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, Dubrovnik, Croatia. 74–85.

Ong, Walter J. 2002 [1982]. *Orality and Literacy. The Technologizing of the Word*. London: Routledge.

Parker, James, and Diana Parker. 1974. "A Tentative Phonology of Baining (Kakat Dialect)." In *Phonologies of four Papua New Guinea Languages*, ed. by Richard Loving, 5–43. (Workpapers in Papua New Guinea Languages 4.) Ukarumpa: Summer Institute of Linguistics. (Cited after Hellwig. 2019.)

Piper, Benjamin, Stephanie Simmons Zuilkowski, Dunston Kwayumba, and Arbogast Oyanga. 2018. "Examining the Secondary Effects of Mother-tongue Literacy Instruction in Kenya: Impacts on Student Learning in English, Kiswahili, and Mathematics." *International Journal of Educational Development* 59: 110–127.

Reetz, Steffen. 2022. Code-Switching among Qaqet-Speaking Adults in Kamanakam, East New Britain Province, Papua New Guinea. PhD thesis, University of Cologne. https://kups.ub.uni-koeln.de/55141/

Riyanto, Edi Dwi, Syahrur Marta Dwisusilo, Shaffarullah Bin Abdul Rahman, and Nadya Afdholy. 2023. "Tengger Communities Creativity and Language Preservation in Text-Speak on the WhatsApp Social Media Platform." *International Journal of Anthropology* 38 (3–4): 263–279.

Schippling, Melanie. 2026. *Speaking of Writing Romani: Language Attitudes, Text Editing, and Variability*. (IMPACT: Studies in Language, Culture and Society 56.) Amsterdam: Benjamins.

Schlüter, Julia, and Katharina Deckert. 2025. "Will AI-Assisted Writing Streamline Non-Dominant National Standard Englishes? A Systematic Case Study of Prepositional Variation." Presentation, Corpus linguistics 2040: Which data, which methods, which models?, Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, 10–11 July 2025.

Schroeder, Christoph. 2014. "Türkische Texte deutsch-türkisch bilingualer Schülerinnen und Schüler in Deutschland." *Zeitschrift für Literaturwissenschaft und Linguistik* 44: 24–43.

Schultze-Berndt, Eva. 2006. "Linguistic Annotation." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 213–252. Berlin, New York: De Gruyter Mouton.

Sebba, Mark. 2009. "Sociolinguistic Approaches to Writing Systems Research." *Writing Systems Research* 1 (1): 35–49.

Seifart, Frank. 2006. "Orthography development." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 275–300. Berlin, New York: De Gruyter Mouton.

Seifart, Frank. 2008. "On the Representativeness of Language Documentations." In *Language Documentation and Description* 5: 60–76. London: SOAS.

Seifart, Frank, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. "Language Documentation Twenty-five Years on." *Language* 94: e324–e345.

Shan, Xinyang, Yuanyuan Xu, Yining Wang, Yin-Shan Lin, Yunshi Bao. 2025. "Cross-Cultural Implications of Large Language Models: An Extended Comparative Analysis." In *HCI International 2024 — Late Breaking Papers. HCII 2024. Lecture Notes in Computer Science*, Vol. 15375, ed. by Adela Coman, Simona Vasilache, Fiona Fui-Hoon Nah, Keng Leng Siau, June Wei, and George Margetis, 106–118. Cham: Springer.

Souag, Lameen. 2010. "Ajami in West Africa." *Afrikanistik Online* 7: 1–11.

Spencer, Piyapath T. 2024. "Documenting Endangered Languages with LangDoc: A Wordlist-Based System and A Case Study on Moklen." In *Proceedings of the Third Workshop on NLP Applications to Field Linguistics*, Bangkok, Thailand. 28–36.

Taylor, Stephen, and Marisa von Fintel. 2016. "Estimating the Impact of Language of Instruction in South African Primary Schools: A Fixed Effects Approach." *Economics of Education Review* 50: 75–89.

Tobar, Edoardo. 2019. "Documenting Online Writing Practices: The Case of Nominal Plural Marking in Zamboanga Chabacano." In *Documentation and Maintenance of Contact Languages from South Asia to East Asia*, ed. by Mário Pinharanda-Nunes, and Hugo Cardoso, 141–173. (Language Documentation and Conservation Special Publication 19.) Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24908

Toninato, Paola. 2014. *Romani Writing. Literacy, Literature and Identity Politics*. (Routledge Research in Literacy 4.) New York: Routledge.

Van Wyk, Jacolynn, and Maria Louise Mostert. 2016. "The Influence of Mother Tongue and Gender on the Acquisition of English (L2). The Case of Afrikaans in Windhoek Schools, Namibia." *Cogent Education* 3 (1), 1–12.

Voßkamp, Patrick. 2010. *Sprechen, um zu schreiben. Mündliche Kommunikation im Lokaljournalismus*. (ESS-KuLtur 2.) Duisburg: Universitätsverlag Rhein-Ruhr.

Walker, Katherine, Birgit Hellwig, and Dagmar Jung. This volume. "Transcription practices and (the documentation of metalinguistic knowledge)." In *The Documentarist Turn: From Observable Linguistic Behaviour to Typological Generalizations*, ed. by Sonja Riesberg, Uta Reinöhl, and Birgit Hellwig. Amsterdam: John Benjamins.

Walsh, Olivia. 2021. "Introduction: In the Shadow of the Standard. Standard Language Ideology and Attitudes towards 'Non-Standard' Varieties and Usages." *Journal of Multilingual and Multicultural Development* 42 (9): 773–782.

Witt, Roxanna-Lorraine. 2021. "Gadjé-Rassismus. Ein analytischer Perspektivwechsel auf Kontinuitäten menschenfeindlicher Ideologien in weißer Kultur und Identität." In *Rassismus. Macht. Vergessen. Von München über den NSU bis Hanau: Symbolische und materielle Kämpfe entlang rechten Terrors*, ed. by Onur Suzan Nobrega, Matthias Quent, and Jonas Zipf, 125–144. Bielefeld: transcript.

Yang, Jianliang, Xiya Zhang, Kai Liang, Yuenan Liu. 2023. "Exploring the Application of Large Language Models in Detecting and Protecting Personally Identifiable Information in Archival Data: A Comprehensive Study." In *Proceedings of the 2023 IEEE International Conference on Big Data (BigData)*, Sorrento, Italy. 2116–2123.

Zhao, Hui, and Hong Liu. 2021. "(Standard) Language Ideology and Regional Putonghua in Chinese Social Media: A View from Weibo." *Journal of Multilingual and Multicultural Development* 42 (9): 882–896.

CHAPTER 12

More diachronic than you think

Historical depth within language documentation corpora and its potential to mitigate two major biases in linguistics

Uta Reinöhl
University of Freiburg

This paper surveys and spotlights historical layers within language documentation corpora, with a particular emphasis on the many languages for which we lack historical written attestation. I highlight the potentials of lesser-prestige and lesser-accessible, often more conservative varieties, as well as the role of oral art. Paying greater attention to existing depth, and intensifying efforts to document lects and genres that point to greater age, will help advance both historical linguistics and typology in significant ways. On the one hand, conservative lects are particularly suitable for the investigation of (micro-)changes given the unusually tight control for other influencing factors. On the other hand, oral art may preserve what is often the only direct evidence of more ancestral forms and structures.

Keywords: language documentation, diachrony, typology, oral tradition, Indo-European bias, Kera'a, Igu

1. Introduction

This paper puts diachrony centre stage in relation to our efforts to document and describe the world's linguistic heritage. I will first make what should be an uncontroversial point, namely that most documentation corpora possess internal diachronic depth even when there is no written historical attestation. I will then lay out an argument that paying greater attention to this diachronic depth, as well as increasing efforts to diversify corpora keeping diachronic depth in mind, will help us alleviate two major biases in historical linguistics and linguistic typology. These are, firstly, the *language bias* of historical linguistics, i.e. the persisting, significant bias towards (Indo-)European and some select other languages

https://doi.org/10.1075/slcs.240.12rei
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

and families. The second bias is what I call the *synchrony bias* of linguistic typology, i.e. the tendency of (especially quantitative) typology to disregard the differing ages of data points between the poles of, roughly, spontaneous dialogue, on the one hand, and (part)ancestral oral art (e.g., rituals, prayers, songs etc.), on the other. This disregard results in a situation where data points emerging from sometimes widely different production conditions and ages are compared without control or consideration, despite their known and fundamental impact on language form. I argue in this paper that documentation corpora can play a pivotal role in reducing both of these biases and thus pave the way towards a more diversified, more nuanced understanding of the human variation space through space as well as time. Building on Himmelmann (1998), this paper, thus, argues for greater awareness of the existing diversity — and for greater diversification — of language documentation corpora not only in the "horizontal" of different types of communicative events, but also in the diachronic "vertical".

The foundation for my argument is that, despite the common perception that languages without written history mostly lend themselves to synchronic, comparative analysis, there is much diachronic depth to explore within most language documentation corpora. Most languages spoken today lack written attestation that is older than a few decades or couple of centuries, and these are the languages that are commonly documented in linguistic fieldwork. This state of affairs, however, does not mean that documentation corpora only consist of contemporary data layers. I will begin in Section 2 with a brief survey of some of the multiple dimensions in which historical depth distinguishes layers in documentation corpora.

I then turn in Section 3 to the first of the two biases this paper addresses: The language bias of historical linguistics. I will briefly survey the current state of the art which continues to be strongly biased towards Indo-European (or just European) languages and languages with deep(er) written attestation. I then build an argument that studies of diachronic micro-changes, as can be detectable within language documentation corpora, can make a significant contribution to diversifying our understanding of possible and probable linguistic changes. Some might object that historical depth is marginal in documentation corpora, and so can only make a modest contribution to historical linguistics. I will show that this is a misunderstanding in two respects: Firstly, diachronic depth holds on multiple dimensions from socio-linguistic variation to genres and structures, and so there is ample ground for historical-linguistic research in every corpus that at least partially lives up to the ideal standards of diversified documentation corpora (Himmelmann 1998). Secondly, and this is the main point, it is micro-changes of the kind observable in language documentation corpora that have particularly great potential to further our understanding of language change. The reason is that we are here presented with an approximation to the historical lab situation

(cp. Reinöhl & Casaretto 2018, Reinöhl, in prep.), where a maximum of factors is controlled for, so that we can draw stronger, more robustly supported insights regarding the factors involved in linguistic change.

I turn to the second bias addressed in this paper in Section 4: the synchrony bias of linguistic typology. The enduring trend towards quantitative approaches across linguistic subdisciplines can be observed in the publication of various large-scale databanks for cross-linguistic comparison and analysis, such as WALS (Dryer & Haspelmath 2013) or Grambank (Skirgård et al. 2023). These types of databanks, however, do not normally control for modality or genre, even while these parameters are commonly known to impact structure. There is even less awareness or attention paid to the different diachronic statuses of data points as correlating with modality and genre. Developing such awareness and control, however, will be crucial in developing linguistic typology towards more robust and reliable generalizations. Using the example of oral art, I will outline how language documentation corpora can help mitigate the synchrony bias that is widespread in, especially larger-scale, linguistic typology.

I will conclude the paper, in Section 5, with a few remarks on practical and ethical consequences and considerations based on the here advocated shift towards taking the diachronic depth of documentation corpora seriously.

2. Diachronic depth within documentation corpora

Historical linguistics normally comes into play for language documentation projects when (i) there are sources at least several decades older than more recently collected material, (ii) in the shape of historical-comparative analysis, comparing documented language varieties with each other where genealogical relations are suspected or (iii) in the form of contact-linguistic research when a history of bi- or multilingualism is attested or plausible. While my focus is here on language documentation corpora more broadly, the points I will raise particularly apply to languages that lack written history of the type mentioned under (i). There is much that one could discuss here, but I will restrict myself to only a few brief remarks on some major diachronic dimensions shaping "synchronic" spoken corpora, with a focus on diachronic clines between and within sociolinguistic varieties.

Between-variety differences obtain, for example, between more conservative and more progressive dialects (e.g., ones exposed to less or to more language contact, for socio-economic or geographical reasons; Trudgill 2011, DeLancey 2014), between older and younger speakers (researchable in a Labovian apparent-time design, while considering lifespan-related change), between female- and male-identifying speakers, between speakers less or more trained in, or exposed to,

reading and writing, between less mobile vs. more mobile speakers etc. It goes without saying that some of these categories may be correlated, and some correlations may hold more broadly (e.g. between age and literacy clines) while others are more society-specific (e.g., whether younger or middle-aged women "lead the change"), and we also know that, at closer inspection, more nuanced, 3rd wave sociolinguistic patterns, emerge. Internal variation is gaining more and more attention in fieldwork linguistics (e.g. Meyerhoff 2019, Lehmann et al. 2025), but is not always viewed from an angle of the more conservative vs. the more progressive, i.e. from a diachronic perspective. Moreover, it tends to be comparatively better-studied languages where scholars have achieved at least a fundamental understanding of socio-linguistic diversity, given limitations of resources and time (Meyerhoff 2019). By contrast, in the race against time facing endangered languages, internal diversity may often not be researchable above and beyond the documentation of, in principle, "only" one variety.[1] Nonetheless, part of the argument of this paper is to call for greater diversification of documentation corpora beyond prestige varieties, where possible.

Within-variety diversity can also be investigated from the perspective of diachronic depth, an endeavour sometimes labelled "internal reconstruction". Here, too, we can build on a wealth of research coming from higher-resource — and some lesser-resource — languages. Besides more obvious time-depth differences between e.g. recent loans and long-attested core vocabulary, other subtler differences can be found between older and younger grams or constructions. More generally, we know that grammatical structures in a given domain may be of different ages, with the phenomenon of "layering" having become part of theorizing about grammaticalization phenomena (Hopper 1991). An example for structural layers shaping an entire grammatical domain comes from Indo-Aryan. The New Indo-Aryan participant-marking system, as in many other Indo-European languages, consists of an old, mostly petrified morphological case system of direct and oblique stems, followed by a younger layer of simple adpositions, and a yet younger layer of secondary, i.e. complex, adpositions (Masica 1991: 230ff; Reinöhl 2016). Another typical domain for internal reconstruction concerns phonology or morphophonology, where e.g. certain allophones are often plausibly older than other younger variants, which arose in the form of sedimentations of contextual influences on, or interactions between, phonemes. Any of these layers and structural subsystems are, of course, not set in stone, but constantly evolving. As Hopper (1987) put it: "always deferred, always in a process, but never arriving" (141).

1. A common situation seems to be that a more conservative variety is more endangered than the progressive (often, prestige) variety, and at the same time further back in the queue for documentation.

Beyond variation-, item- and structure-related historical depth there is another important domain in which we find different diachronic layers: genre. I will postpone general remarks to Section 4, where I discuss a diachronic perspective on genre diversity in more detail.

3. The language bias of historical linguistics

I begin with a brief survey of the persisting bias of historical-linguistic research towards, in particular, (Indo-)European languages (3.1). I then turn to the potentials of historical micro-changes, as potentially manifest in sufficiently differentiated language documentation corpora, to considerably diversify historical linguistics. I argue that it is the historical lab approximation of tightly controlled changes, i.e. where social and system-related factors are maximally held constant, that makes historical depth in documentation corpora an unusually well-suited research field to extend our understanding of what are possible and probable language changes (3.2).

3.1 The state of the art

It is well known that historical linguistics is biased towards languages and language families with a written history, especially deeper and/or historically less uninterrupted written history. Additional biases such as Eurocentrism and a bias for the languages that became lingua francas or are/were the languages of larger religions can be added. These biases remain active as is confirmed by a survey of the last hundred papers published in *Diachronica* and *Folia Linguistica Historica* (Table 1).[2]

What is interesting to observe is that this distortion seems to be only rarely addressed as a limiting factor for historical-linguistic description and theorizing (but see remarks by Epps, Daniels, and Tan in Bowern et al. 2024 and by Reinöhl in Bowern et al. 2025). A comparison with the situation in linguistic typology can be revealing in this regard. Typology has, to a certain extent, shifted in the last decades from a strong belief in ("statistical") linguistic universals in the Greenbergian mould to a greater awareness of the fact that living languages today are all but a fraction of languages that have been spoken in human history, and so

2. Given what I take to be an uncontroversial bias towards (Indo-)European and few other languages and families, I abstain from presenting this survey in a more statistically refined way, which might weight studies by number of languages per family, speaker numbers, biases towards individual languages within families etc.

Table 1. Top ten language families studied in the 100 most recent (before May 2025) papers published in *Diachronica* and *Folia Linguistica Historica*

Diachronica		**Folia Linguistica Historica**	
Indo-European	42	Indo-European	78
Trans-Himalayan (Sino-Tibetan)	15	Trans Himalayan (Sino-Tibetan)	9
Austronesian	9	Afro-Asiatic	6
Uralic	5	Austronesian	4
Afro-Asiatic	4	Japonic	3
Niger-Congo	4	Nakh-Daghestanian	3
Basque	3	Niger-Congo	3
Japonic	3	Penutian	3
Tungusic	3	Uralic	2
Turkic	3	Arnhem	1

sampling cannot be easily generalized.[3] Recurrent patterns that do exist hardly ever do so without exception, and represent practical design solutions for cognitive and communicative tasks and goals (Evans & Levinson 2009). Moreover, many, if not most, modern languages show some degree of contact influence from a small number of languages, in particular the languages of European colonialism, thus increasing distortion in the last few centuries (see e.g. Moran et al. 2021). Nonetheless, diversity remains significant. As Evans & Levinson (2009) put it, "language universals are a myth" on all levels of linguistic structure. Current research suggests that the more robust universals are those relating to language use, e.g., mechanisms of turn taking (Stivers et al. 2009, Levinson 2006), rather than to structural patterns. In linguistic typology, then, there seems to be a general recognition that each new language that is described may, or is even likely to, teach us something new about linguistic diversity. While the Indo-European bias persists in linguistic typology, too, it is significantly reduced relative to historical-linguistic research (Table 2).

It is clear that historical-linguistic work lags behind synchronically oriented work in terms of language diversity. However, while "typological" may not be in the name of this subdiscipline, I don't doubt that historical linguistics is, in principle, no less ambitious than linguistic typology to map out the human linguis-

3. There are multiple additional reasons why many linguists seem to have grown more sceptic of representativeness and universality claims. An obvious one is internal variation, where e.g. coding a language for just one (dominant) word order hardly ever does justice to actual complexity.

Table 2. Top ten language families studied in the 100 most recent (May 2025) papers published in *Linguistic Typology* and *Language Typology and Universals (STUF)*[*]

Linguistic Typology		**Language Typology and Universals (STUF)**	
[studies covering more than 10 language families]	24	Indo-European	58
Indo-European	12	Afro-Asiatic	26
Sino-Tibetan (Trans Himalayan)	8	Niger-Congo	20
Austronesian	6	Austronesian	19
Niger-Congo	4	*[studies covering more than 10 language families]*	16
Afro-Asiatic	3	Uralic	16
Austroasiatic	3	Sino-Tibetan (Trans Himalayan)	15
Basque	3	Kra-Dai	11
Uralic	3	Trans-New Guinea languages	10
Japonic	2	Japonic	6

* Papers investigating more than ten language families are assigned to the category "studies covering more than 10 language families", and the language families included were not additionally counted individually. Note that this category is among the top ten by frequency only for these two typology journals, but not for the historical journals in Table 1. While the number of papers investigating Indo-European languages in *STUF* is higher than in *Diachronica*, some other language families also have significant coverage in *STUF* in contrast to the comparatively low numbers for other language families in either of the historical journals.

tic variation space — in this case through time. I am also certain that most, if not all, historical linguists would agree that it is a desirable aim to study more languages from a historical perspective. However, I suspect that the primary reason for this is the value put on description — simply the widening of how much we know about the languages of the world — rather than the expectation that we will have to, perhaps even fundamentally, revise our theoretical models or adapt our expectations with regard to possible or probable changes.

I suspect that this situation — where description may be valued more for itself and less due to expectations of (fundamentally) new historical-typological insights — is connected to some influential thought traditions in historical linguistics.[4] For example, there is a long tradition, dating back to the Neogrammar-

4. Another reason is likely the very absence of historical written attestation, which might lead one to think that older forms can only be reconstructed, not investigated. I show in this paper that this is not so, but that, on the contrary, most language documentation corpora will attest to forms of at least some historical depth.

ians, to regard certain sound changes as particularly likely. A consonant cluster is exceedingly more likely to lose, rather than gain, a segment, and intervocalic position likely leads to lenition rather than strengthening. We also expect reduction and loss at the right edge of phonological words, but not at their left edge (Fougeron and Keating 1997, Keating et al. 2003). In morpho-syntax, too, we are used to expecting the recurrence of certain grammaticalization pathways. In fact, the very metaphor of "paths" or "trajectories" suggests that grammatical change takes place in similar, rather than widely diverse, ways. Bybee & Dahl (1989) argue for the existence of a "small set of cross-linguistic gram-types, identifiable by their semantic foci and associated with typical means of expression" (1989:52), a spirit running through much of grammaticalization literature. And indeed, the evidence from "go-futures", body-part terms developing into adpositions and case markers, or indefinite articles grammaticalizing from the numeral 'one' underscores that these expectations are certainly not without foundation. Many similar cases could be added.

However, over time many authors have pointed out that some of our assumptions of what is common and expected may not hold up if samples were more diversified. One example is Bisang (2009) who argued that grammaticalization pathways follow different trajectories across Southeast and East Asian languages, with different gram profiles — in meaning and in form — in comparison to Indo-European pathways of change. A key reason identified by Bisang lies in the different phonological conditions on change. Stress-based systems, common in Indo-European, favour right-edge reduction (and thus the cline from word to clitic and affix). Systems lacking stress, and/or where words are overwhelmingly monosyllabic, show different trajectories, where grams tend to remain phonologically independent, or maximally reach a clitic stage. Other authors have studied rare changes in phonology (e.g. Blevins 2007; contributions in Kuznetsova et al. 2025) or morphology, some of which question fundamental theoretical or typological assumptions (e.g., the basic status of the CV syllable, and left-edge prominence, in Blevins 2007). Such findings do not come as a surprise: Again, according to Evans & Levinson (2009) and the rich literature they survey, there are few things which do not occur (while there are reasons for why certain phenomena are frequent and others rare, even while we might need to update which are which, too, if samples become more diverse). If that is so, then, from a historical viewpoint, all of these patterns and phenomena must have arisen somehow, and thus await a historical-linguistic account. And there are probably many more patterns and phenomena that still await linguistic description and typological awareness.

In the realm of low-resource languages, historical accounts have traditionally leaned on written evidence or comparison with other varieties (and possibly recon-

struction). In the next section, I will illustrate how corpus-internal depth, too, can make a substantial contribution to historical-linguistic research, in ways that have special potential for empirical insight and theoretical elaboration given the tight control for potential influencing factors.

3.2 The historical lab approach in language documentation

As Meyerhoff (2019) observes, language documentation is necessarily confronted with variation, but typically with a view to select and/or generalize from it, rather than in order to take full(er) stock of it. This is a natural by-product of limited resources in the face of un(der)explored varieties: "Linguists working in the field of language documentation, too, are necessarily engaged with language variation, but from a somewhat different perspective. Instead of starting from a well-understood and well-documented language grammar, documentary linguists are using their skills to (first) extrapolate away from the inter-individual variation that characterizes all speakers and every speech community in order to adequately describe the structure of a lesser-known or poorly-documented language." (Meyerhoff 2019: 230) There are of course numerous examples where fieldworker-grammarians have taken considerable note of language variation (examples include Evans's (2003) pan-dialectal grammar of Bininj-Gunwok or Hualde & de Urbina's (2003) grammar of Basque). In other cases, which can now be considered best practice, fieldworkers are explicit about the specific sub-variety they research even while resource limitations may preclude wider coverage.

Rather than discussing the interplay of fieldwork and variation more generally, I will focus on what selectiveness observed by Meyerhoff means with regard to diachronic dimensions. Trying to come to terms with variation often results in the linguist's orientation towards a prestige variety or variety spoken in more easily accessible rather than more remote locations, which, in turn, comes with correlations regarding prestige and/or infrastructural connectedness. (Often, speakers themselves will express their stances on the topic of which variety should be privileged in documentation, in interactions with fieldworkers. Such prioritizations tend to be embedded in complex local societal-political dynamics, which a fieldworker may not have a full understanding of, but will nevertheless need to engage with.) From a linguistic viewpoint, however, there is evidence that prestige varieties differ from non-prestige or "dialectal" varieties in generalizable ways (Trudgill 2008).[5] For example, it is one of the more intensively researched correlates of typological diversity that language contact and especially L2 learning

5. Such differences may be exacerbated if a variety becomes the model for a written standard, whereas others remain essentially unwritten.

leads to morphological simplification (Lupyan & Dale 2010) and so a field linguist's selection of one, possibly better accessible and more contact-influenced variety over another is not without repercussions.[6] From a historical angle, however, non-prestige and lesser-contact varieties are known to be more likely to preserve inherited features. Examples are aplenty, including not only endangered languages but also many languages spoken in Europe (Auer 2018), even while individual contact situations may not necessarily lead to simplification (Meakins et al. 2019). While resource limitations are serious constraining factors, I would like to make a case for including non-prestige and lesser-contact varieties, if at all possible, for the dual reasons that these may be even more endangered — and thus their documentation more time-pressing — and that they may also hold invaluable insight for historical-linguistic research.

An example comes from the ongoing loss of word-initial consonants in CV syllables in Kera'a (or Idu (Mishmi); idum1241; clk), a Trans-Himalayan language predominantly spoken in Arunachal Pradesh, India (Reinöhl 2022, Reinöhl & Peck, in prep.). Word-initial consonant loss is theoretically and typologically unexpected; it stands in conflict with the assumed prominent status of word-initial consonants and, arguably, also with the assumed naturalness of CV syllables. This is (so far) matched empirically; outside of Australia, word-initial consonant loss is rarely attested, typically only affects one phoneme and, quite often, evidence for loss has only been indirectly adduced based on reconstructions (Blevins 2007, Reinöhl et al., under review).

Word-initial consonant loss in Kera'a affects the prestige dialect Midu, which means that a language documentation project focusing only on this variety would not catch it in the first place (as it would not present as loss, but simply in the form of vowel-initial words). In the lesser-prestige dialect Mithu, which is spoken in the harder-to-access, steeper mountain ranges away from the plains in upper Lower Dibang Valley, many words retain one of a range of initial consonants.[7] Initial loss affects the consonants before non-back, non-nasalized vowels in minimally disyllabic words across (mostly content-word) lexemes. The following three Tables (3–5) from Reinöhl et al. (under review) illustrate the affected consonants in their environments and offer examples for each attested (C)V syllable.

6. Of course, L2 learning is only one correlate of prestige, and may not play a role for languages spoken by more marginalized, or smaller societies.

7. Mithu homeland is not only harder to access, with lower speaker numbers, but its marginalized status within Kera'a society has been attested since the 19th century. The former Assamese-derived term for the Mithu, adopted by the British in colonial times, was *Bebejiya* 'outcast' (Reinöhl 2022).

Table 3. Consonant inventory of Kera'a with consonants affected by word-initial consonant loss indicated in bold and shaded

	Bilabial (unrounded)	**Bilabial (rounded)**	**Alveo-dental**	**Palatal**	**Velar**	**Glottal**
plosives	**p** p^h ~ f <ph> **b**	**p^w** <pw> **ph^w** <phw> **b^w** <bw>	**t** t^h <th> d		**k** k^h <kh> g	
nasal	**m**	**m^w** <mw>	n		ŋ <ng>	
fricative			s	ʃ <sh>		**h**
affricates			ʦ <ts> ʣ ~ z <z>	ʧ <ch> ʧʰ <chh> ʤ <j>		
approximant		w	**r**	**j** <y>		
lat. approx.			l			

Table 4. Kera'a vowels, with Vs in word-initial CV syllables affected by word-initial consonant loss in bold and shaded

i , ĩ		u, ũ
e , ẽ	o, õ	
a , ã		

For Australian languages, scholars have suggested a variety of explanations for word-initial consonant loss which is attested in at least several dozen languages (Blevins 2001). The seemingly most wide-spread proposal relates to stress, where a stress shift from the first to a later syllable is assumed to have somehow weakened the word-initial position. Other proposals rely on "phonetically weak" consonants disappearing first, including glides and non-strident fricatives (Blevins 2001). The third cluster of proposals refers to position-related phenomena, since the loss appears to, in some languages, affect words that are particularly frequent in utterance-initial position (Alpher 1976, Blevins 2001).

The tight parallels between Midu and Mithu phonetics and phonology allows us to exclude all three of these types of explanations. As shown in Culhane et al. (2025), there is no evidence for the existence of lexical stress in Kera'a. Rather, potential phonetic correlates such as pitch and duration are indicative of tone, with Kera'a showing a typologically and areally unusually complex tone system

Table 5. Examples for each attested consonant-vowel combination

CV shape	Mithu	Midu	Translation
/pi/	/**pi**tʃiko/	/**i**tʃiko/	'bark'
/pa/	/**pa**mbõhã/	/**a**mbõhã/	'kneecap'
/bi/	/**bi**li/	/**i**li/	'pig'
/ba/	/**ba**ha/	/**a**ha/	'to get lost'
/mi/	/**mi**ku/	/**i**ku/	'dog'
/me/	/**me**to/	/**e**to/	'chicken'
/ma/	/**ma**tī/	/**a**tī/	'village'
/p^{w}i/	/**p^{w}i**di/	/**i**di/	'pure, without blemish'
/p^{w}e/	/**p^{w}e**tre/	/**e**tre/	'comb'
/b^{w}e/	/**b^{w}e**tambõ/	/**e**tambõ/	'Russell's viper'
/m^{w}i/	/**m^{w}i**ri/	/**i**ri/	'to threaten'
/m^{w}e/	/**m^{w}e**se/	/**e**se/	'to kill'
/ta/	/**ta**hru/	/**a**hru/	'to ask'
/ka/	/**ka**jo/	/**a**jo/	'rain'
/ra/	/**ra**tu/	/**a**tu/	'up, above'
/ja/	/**ja**hũ/	/**a**hũ/	'bear'
/ha/	/**ha**p^{h}u/	/**a**p^{h}u/	'cheek'

(Reinöhl & Peck, in prep.).[8] The second explanation also cannot account for the loss in Midu given that both phonetically weak and strong consonants are affected, with no evidence of weak consonants leading the change (Reinöhl et al., under review). (On the contrary, some phonetically strong consonants in Blevins' (2001) sense, such as /k/, are among those most frequently affected, on a lemma and token basis.) The third explanation of position-related effects can also be excluded, with affected words taking a variety of positions in both Mithu and Midu, which is not surprising given the wide spread of the phenomenon over word classes (Reinöhl et al., under review). Instead, the most promising lead is that the change began via lenition of an initial /t/ to a glide /y/ and onwards to loss, which then analogically spread to other CV contexts and lemmas (Reinöhl et al., under review, for details).

As mentioned above, most accounts of word-initial consonant loss in languages outside of Australia are based on reconstructions, and this is also true for some Australian languages. For some of the explanations mentioned above, scholars have had to make additional assumptions in order to fit the data with the

8. Kera'a's tone system displays significant sociolinguistic variation of what appears to be a 5-way level-tone system in some, but possibly not all, varieties. Given the heterogeneity of tonal realization across varieties, and the ongoing status of research, I abstain from marking tone in this paper, but see Reinöhl & Peck (in prep.) for data and discussion.

explanation. For example, for some Australian languages, not one, but two stress shifts need to be assumed, from the first to the second syllable and back again to the first syllable, in order to fit the data with a stress-based explanation (Blevins 2001). With Midu, we can work with the attested control case — i.e. Mithu — allowing us to be much more certain of what factors may, or may not, have played a role in the loss. (Additional data from Kera'a's sister language, the phonologically more conservative Tawrã, confirm the loss, see Reinöhl 2022.) Thus, internally diversified language documentation corpora may allow investigating changes that are not only maximally controlled — with an optimum of factors held constant given the proximity of varieties — but also with all varieties in question attested.

Of course, numerous case studies could be named here where authors have, in similar ways, researched variation and advanced our historical-linguistic understanding based on it. And it is this type of internal variation which — as it both expands our knowledge of endangered and lesser-studied languages as well as possible changes — I spotlight here, so that documented diversity of this kind shall receive more attention in historical-linguistic theorizing. A second concern is to intensify fieldwork into lesser-prestige and lesser-contact varieties, where possible in view of resource-related as well as local societal factors.

4. The synchrony bias of linguistic typology

I now turn to a second bias that, as I argue, may be alleviated if we start paying greater attention to the existing diachronic depth within documentation corpora. I will first comment on the widespread lack of metadata or control relating to the genres and diachronic depths of data points used in (larger-scale) typological comparison (4.1). I will then focus on what is perhaps the most important genre for research into historical depth in the absence of written attestation: oral art (4.2).

4.1 Apples and oranges

Language descriptions based on documentation corpora form a primary source for cross-linguistic databanks such as WALS or Grambank.[9] Due to the scale and intent of these databases, a majority of sources derive from understudied, often endangered languages with little if any historical written attestation.[10] In these and

9. https://wals.info/ (last access, 08/06/2025); https://grambank.clld.org/ (last access, 08/06/2025).

10. Even where such attestation exists, it tends not to be included in language documentation corpora, or the derivative large-scale databanks, for that matter.

similar databanks, and the numerous publications that have come out over recent years based on them, data points from different languages tend to be directly compared, rather than to be differentiated with regard to genre or other biasing factors. At closer look, however, the primary sources feeding databanks of this type often consist of heterogeneous collections of linguistic material — oral or written, more spontaneous or more planned, more casual or more formal and/or ritualized, and of different diachronic depths.

We know from the long research tradition of orality and literacy research that these different data types not only differ significantly in structure (Chafe 1994, Ochs 1979, Himmelmann 1998: 178–180), but that they also show historical layering (Ong 1982, Gaenszle 2018). This is particularly true for oral art, which Himmelmann (1998) identifies as one of the five main data types that a comprehensive documentation project should include ("ritual") — provided that oral art traditions survive in the community in question (see Figure 1). Oral art refers to poetically structured verbal performances of spiritual, mythological or traditional-moral character, such as the early Sanskrit Vedas, the Old English poem Beowulf, or the Igu ritual recitations of the Kera'a (see below). Traditional speech practices of this type exist through the oral transmission of long(er) stretches of discourse, sometimes many hours long, in largely formally faithful form (if not fully faithful and/or with in-built variation) (Ong 1982, Kelly 2015). Also, genres not quite at, but close to, the pole of plannedness (Himmelmann 1998) rely on traditional transmission of larger discourse units. This includes Himmelmann's category of "narratives", such as folk stories, a (sub-)genre ubiquitous in language documentation corpora.

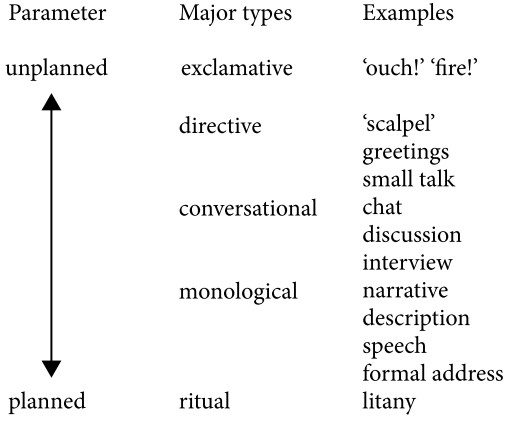


Figure 1. Types of communicative events according to spontaneity (from Himmelmann 1998: 180)

Table 6 surveys the first 15 openly or readily accessible sources fed into Grambank, in alphabetical order (n/a for not applicable, n/m for not mentioned).[11] In addition to modality and the presence of writing, other parameters bearing on the type of speech event are captured. For our purposes, more routinized and less spontaneous content, about more publicly shared topics (even where privately performed, such as prayers), are plausible suspects for showing at least partially deeper diachronic depth.[12]

This selective survey of a fraction of sources underlying Grambank illustrates the significant diversity of data types. These, however, are directly compared in large-scale comparative studies drawing on this or similar databanks. Note that the information presented in Table 6 was retrieved from the original sources, rather than it being made available by Grambank — i.e. it is somewhat hard to come by even where it exists. The numerous gaps in retrievable metadata, even from primary sources, are also noteworthy. What seems clear is that large-scale studies drawing on these and similar datasets compare apples with oranges in multiple ways, with regard to modality, production conditions (spontaneous, elicited, semi-spontaneous, stimulus-based etc.), genre, and — what is the focus of this paper — the diachronic status of data points.

4.2 Oral art and historical depth

I will, in the remainder, focus on the genre that most clearly tends to conserve (partially) older historical layers: oral art. Building on Ochs (1979), Himmelmann (1998) argues that spontaneity or plannedness is the primary dimension along which speech events, and, by consequence, language structures, differ. With oral art at the pole of non-spontaneous speech production, this genre takes a special place in language documentation as it delimits the diversity space with regard to planning degree.[13] I argue that it also takes a special place relating to its conser-

11. A longer data set surveying the first 100 sources of Grambank, including more parameters as well as original sources, can be found in this OSF archive: https://osf.io/k6t35/.

12. But note that some parameters show substantial cross-cultural diversity. For example, multi-party, participatory narration, where reference is co-constructed, has been described for mythological story-telling in several Australian societies (Hill 2018). This contrasts with the association of the monologue with traditional western (post-Socratic) rhetoric and verbal performance.

13. "Oral art" is often used interchangeably with "oral tradition", i.e. orally transmitted, poetically structured language, which is recited or performed in ritualistic contexts (see e.g. contributions in the journal *Oral Tradition*). Poetic forms may vary widely, but are typically characterized by heavy use of repetition and repetition-with-variation patterns in meaning and/or form (parallelism). These arise in connection to ritualization, mnemotechnical ease

vation of at least partially more ancestral layers and structures (Gaenszle 2018, Ong 1982). As Gaenszle writes on oral art in the Himalayas: "There is no doubt that the basic features have been stable for many generations, and thus the linguistic repertoire often contains what can be described as archaisms, e. g. lexical forms that have retained older morphological characteristics. But here again a word of caution is called for. Usually the oral recitations are "compositions-in-performance", i.e. not fixed like written texts, but built from stable formulas. And thus, they are always in a somewhat fluid state, they can be and have been subject to some degree of change." (Gaenszle 2018: 7–8). Disentangling genre effects from chronology effects for such compositions-in-performance presents the key challenge in assessing the extent to which the speech of an oral tradition attests to diachronic depth.

A case in point comes from a comparison of casual speech in contrast to oral art in Kera'a. The Kera'a refer to the oral tradition of their shamanic recitations as *Igu*, a word meaning 'shaman', 'shamanic ritual' as well as 'language of shamanic ritual'. Igu remains actively practiced and more vital than many other shamanic traditions in the (Eastern) Himalayas with, so far, no syncretization evidence with regard to Hindu or Christian influence (see Reinöhl 2024, Reinöhl to appear, Reinöhl et al. 2025).

Igu clause (or rather, line) structure can be characterized as on the whole less integrated relative to clause structure in (casual) Kera'a speech. Many lines in Igu recitations do not display a predicate-argument structure, but rather what appears to be two or three independent speech acts, often in performed direct speech addresses of the shaman towards different types of threatening spirits (Example 1). This phenomenon is accompanied by a far-reaching absence of adpositional marking, whether on arguments or adjuncts, even in those cases where we *can* detect clausal structure, i.e. where the nouns or noun phrases func-

---

and/or aesthetic elaboration (e.g., Jakobson 1966; Fox (ed.) 1988; Kelly 2015). I here use the term oral art for clarity, as oral tradition may also be understood in a broader way, such as to include Himmelmann's (1998) category of (traditional) "narratives". Speech events falling into this broader category may be traditionally repeated in a community, but in formally more varied, and less poetically structured, ways. Oral art forms may be more interactive or more monologic, and much could be said about their diversity. The important point is that there is, to a certain extent, a repetition of largely the same content in largely the same form, and for a recognisably stable societal purpose. "Same" rarely means identical. Repetition faithfulness varies between oral traditions, and, overwhelmingly, repetition-with-variation is the norm, whether for aesthetic or social purposes (where, e.g., a singer or shaman reaches for a distinct style), or for reasons of memorization constraints. However, verbatim transmission of extensive, multi-hour performances is also known, such as, famously, in the case of the early Hindu oral traditions (available at https://vedaweb.uni-koeln.de/, Reinöhl et al. 2018, Kölligan et al, ongoing).

Table 6. A survey of orality/literacy parameters of sources included in Grambank

	Spoken/ written data	**Writing practiced/ not practiced**	**Less/more spontaneous**	**Familiar/ new topic**	**Familiar/ unfam. interlocutor**	**Formal/ inform.**	**Monologue/ dialogue or multi-party**
Abenlen Ayta	spoken	n/m	both	n/m	n/a	n/m	monologue
Abkhaz	n/m	practiced	n/m	n/m	n/m	n/m	n/m
Abu' Arapesh	spoken	practiced, mainly Tok Pisin	n/m	n/m	n/m	n/m	n/m
Abui	spoken/written	practiced	both	both	n/m	both, more informal	both, more monologue
Abun	spoken/written	practiced	both	n/m	n/m	both	both
Aché	spoken	n/m	spontaneous	familiar?	n/m	informal	more monologue
Acehnese	spoken	n/m	spontaneous	familiar	familiar	informal	both
Achuar-Shiwiar	spoken	n/m	both	familiar?	n/m	both	both
Adang	spoken	unclear (written by native speaker and says "their cultural heritage [...] is only transmitted orally from one generation to the next through the language"					n/m
Adonara	spoken	practiced	spontaneous	familiar	familiar rather	informal	more dialogue
Aghu	written	practiced	n/m	n/m	n/m	n/m	n/m
Agod-Ende-Kawam	spoken	n/m	prompted	n/m	n/m	n/m	more monologue, "narrative"
Aguaruna	mostly spoken, some procedural texts	practiced	spontaneous + some elicitation	familiar	n/m, but more monologue	more informal	mostly monologue
Ahin-Kayapa Kalanguya	both	practiced	n/m	n/m	n/m	both, not clear	more monologue
Ahom	written (chronicles), recording of read texts	practiced	n/m	n/m	n/m	n/m	n/m

tion as arguments to a verb (Example 2). For comparison, see an example of a typical (moderately spontaneous) Kera'a clause which displays a transparent predicate-argument structure with almost all participants overtly case-marked (Example 3).

(1) Igu (Trans-Himalayan)
a. *alodju atogi dji-mi-ma*
holy_turmeric tong sit-NEG-POL
'(Shaman says: I here have protective) turmeric! (I here have protective) tongs! Please stay away!'
b. *abrapo alone dji-mi-ma*
surname_of_turmeric short_tong sit-NEG-POL
'(Shaman says: I here have protective) turmeric! (I here have protective) short tongs! Please stay away!' [ayi_a_8:51]

(2) Igu (Trans-Himalayan)
mudu Atuli thru-zu-za-ne …
sky A. run-?-come_down-CONV
'Atuli came running from the sky, …' [kaliwu_9:41]

(3) Kera'a (Trans-Himalayan)
imu / (0.7) *khengge=ni / a(h)e alochi kolombo=ma / naashpati-bõ*
person one=ERG that path middle=LOC pear-tree.CL
dji-(g)a-doma a(h)e ma / dophra phrẽ-de-ne atiri shu-de-ne
sit-CONT-NOMLZ DEM LOC ladder lean-stand-CONV above climb-reach-CONV
'one person, in the middle of that path, where a peartree stands, a ladder leaning, (he) climbs up' [keraa-20160226-UR-01-T-pear_story_00:30]

I have argued elsewhere that the looseness of Igu line structure can, in part, be interpreted as a genre effect where imagined interactions, including direct speech, between the shaman and diverse types of spirits, are ubiquitous (Reinöhl et al. 2025). However, I also believe that the dearth of participant marking even where arguments do satisfy a finite predicate's semantic roles may be indicative of a relatively recent spread of adpositional marking. This is further supported by differences between Midu as the more progressive and Mithu as the more conservative dialect. While Midu displays overt adpositional marking as a default, zero marking is more common in Mithu (Reinöhl & Peck, in prep.). Rather than pursuing this point here in more detail I would like to make the more general, obvious, point: It makes a difference whether or not a databank representation of Kera'a is based on Kera'a casual speech, only or predominantly, or whether a documenta-

tion corpus feeding such a representation would include (or even focus on) Igu oral tradition data.[14]

The disentanglement of history from genre poses a significant challenge, and remains a major point of contention in prominent historical-linguistic debates, but has hardly penetrated typological research considerations. For example, discussions around word order flexibility and word order defaults in early Indo-European languages have long been swaying back and forth between scholars emphasizing the genre specificities of e.g. the Rigvedic hymns in contrast to other scholars putting less emphasis on genre and arguing for actual syntactic change between older poetically structured and younger prose sources (Hock 1997). A promising way out is to approximate control for genre as much as possible, e.g. through comparing oral-tradition speech events with e.g. the (free(er)) rendering of a traditional narrative, and thirdly to even less planned genres. If the traditional narrative resembles, in some structural aspects, those of oral-art speech events, while showing robust differences from more casual speech patterns, a contribution of genre would seem likely. On the other hand, if there are more substantial differences between oral-art and traditional-narrative structures, in comparison with greater similarities between the latter and other less planned genres, differences setting oral art apart may indeed relate to change.

While oral art does not survive in every society speaking an endangered or understudied language, I already mentioned that another, much more widely attested genre is also relatively close to the more planned, less spontaneous, scale of Himmelmann (1998): the folkstory. This genre, in turn, inherently involves traditional transmission, albeit with a lower degree of, especially, formal faithfulness. I believe that the time is ripe to enrich typological databanks with metadata that allows testing for more spontaneous in contrast to more planned speech, and for orality/literacy variation more broadly (where the modality-based binary is complemented by a conceptual continuum from the more prototypical oral to the more prototypical literate, Koch & Oesterreicher 1985).

In addition to comparing more spontaneous with more planned speech events, it is especially the role of writing that remains barely studied with regard to its impact on cross-linguistic diversity. Anthropologically grounded research into so-called primary vs. secondary orality (Ong 1982) — i.e. orality in societies without and with writing respectively — is replete with stark claims around the

14. One reviewer raises the possibility that the dearth of adpositional marking is fully genre-dependent, rather than indicating change. While this possibility cannot, at this point, be ruled out, the cline of adpositional marking from Igu (rare) to Mithu (optional, with differential marking) and Midu (near-obligatory for some semantic roles) is suggestive of change, with Mithu also in other respects more conservative than Midu.

impact of writing onto language structure. While there are studies arguing that, e.g., the rise of literacy leads to more syntactic embedding (e.g. Höder 2010, Maas 2010), these and related orality/literacy claims have hardly been tested on a larger cross-linguistic scale (one partial exception is Becker et al. 2023). While the literature is biased towards focusing on languages with historical written attestation in historical linguistics and on languages without historical written attestation in fieldwork and larger-scale data bases, the time is ripe to integrate insights into the impact of modality-based and conceptual orality and literacy dimensions and their diachronic correlates with typology.

5. Conclusion

This paper argues that documentation corpora display a diachronic depth which is often overlooked, while it has significant potential to reduce two persistent and potent biases in linguistics. I have shown, first, that historical linguistics retains a strong focus on Indo-European languages, naturally restricting our knowledge base when it comes to possible and probable types of language change. Here, micro-changes attested within language documentation corpora — e.g. between more conservative and more progressive varieties — have a special potential to extend and diversify our understanding of change in human language given the unusually tight control of the surrounding language system and patterns of language use. The second bias concerns the widespread disregard of large-scale typological comparison for the nature of data points with regard to their modalities, production conditions and genres. While there seems to be some awareness of these types of influencing factors on language structure, they are hardly ever controlled, let alone annotated for. I have thus revisited Himmelmann's (1998) spontaneity/plannedness cline of speech events through a diachronic lens, arguing that oral art, in particular, as well as other genres that involve the traditional transmission of larger discourse units, tend to contain (part)ancestral remnants (while not being closed off to change). In other cases, formal differences between e.g. oral art and more spontaneous speech events may be due — in part or in full — to genre differences, rather than indicative of linguistic change. In our endeavour towards a deeper understanding of the human linguistic variation space, it appears incumbent on us to start considering and controlling for these factors, including in the form of metadata enrichment of data in larger-scale data banks.

I have also advocated in this paper for more attention to be paid to more conservative lects and genres that bear special potential to advance both historical linguistics and linguistic typology. It is, however, one thing to push for certain adjustments in research practice, and it is another how a speech community may

position itself towards such suggested adjustments. The two data types that are in the focus of this paper — conservative dialects and oral art traditions — appear to occupy two rather different positions in this regard. I will therefore conclude with a few remarks relating to some political and ethical dimensions of what I have argued for in this paper.

In his seminal paper, Himmelmann (1998) makes the following remark: "I presume without further discussion that the interests and rights of contributors and the speech community should take precedence over scientific interests." (1998: 172). I believe that there are, indeed, many cases where a fieldworker's actions do align with a community's interests, e.g. when the fieldworker prioritizes a prestige variety that is also prioritized for documentation by (the majority of) a speech community. (At the same time, the fieldworker may want to make sure, if at all feasible, that the speech forms of minoritized sub-communities are also documented. In a way, following the community's lead in terms of the choice of variety to be documented may feel treacherous, as it recreates a society's societal asymmetries.)

If community interests took exclusive precedent, however, I believe that my own research with the Kera'a would almost only focus on Igu shamanic rituals,[15] with an additional, limited amount of traditional story-telling and oral history in Kera'a, most likely by select, socially prominent members of the community. I also venture the guess that this is not an unusual situation, namely that the actual interests and priorities of speech communities diverge, in part fundamentally, from what a fieldworker may wish to focus on. (Of course, interests and prioritizations can be discussed and explained, and adjusted together with members of the speech community.)

This commonly found, heightened interested in oral art is, however, at odds with the long trend in linguistics from a focus on written, redacted language in the 19th century to a focus on speech in the 20th century. The more recent focus on speech events that are as spontaneous, unprompted and even unobserved as pos-

15. A personal anecdote may be of interest here. In my own fieldwork I have very early on begun to work on the shamanic language Igu, the language of oral tradition among the Kera'a. At a conference presentation on Igu, I was asked by a fellow fieldworker with a shared regional focus in the Himalayas, why and how I went about documenting both, the genres of everyday communication and those of oral art at the same time. The researcher said, more or less verbatim: "Many of us others document the language of everyday communication for 30 years before we think we know enough about it so that we can turn to oral art." Unfortunately, this stance can easily lead to the loss of an oral tradition before we feel "ready for it". Even if an oral tradition is persistent enough to wait that time, practitioners may lose the level of competence that used to be common in the olden days. Of course, there is also no intrinsic reasons for any unidirectionality in how research into either genre, register or variety can inform the other.

sible, takes us even further away to the opposite pole from oral art on Himmelmann's plannedness/spontaneity cline. We can here identify a third bias, which I might term the "spontaneity bias" of contemporary documentary linguistics.[16] As Himmelmann puts it: "Among the targets of such criticism [of traditional language description, UR] are its abstract and ahistorical conception of the speech community as a homogeneous body, its neglect truly to confront the complexities of spoken language (rather than reducing spoken language to "language as it may be written down") and the concept of a language as an overall coherent system." (1998: 164) Thus, spontaneous speech as found in (casual, non-scripted) conversations today takes a central role in many language documentation efforts, as the most valued data type in some respects, and in more ways than one the most complicated and challenging data type, too, rendering its research particularly worthwhile to pursue. As Himmelmann (1998) observes, conversational language use is also the type of genre that speakers engage in most, and so its importance in language documentation corpora mirrors its role in speakers' lives. Last but not least, spoken language research, and particularly spontaneous speech research, has a lot to make up for given a past where it was almost entirely disregarded.

There is no doubt in my mind that some of the most linguistically exciting research in language documentation lies in the realm of interactional, multimodal linguistics, but the fact that we have made progress at the spontaneous end of the Ochs and Himmelmann dimensions should not result in a situation where the other end of the spontaneity spectrum is neglected, particularly when it might take such a central position in speakers' cultural imaginations. While it is true that moderately planned genres — like the above-mentioned folk-story — remain omnipresent in documentation corpora, this is not so for the genre at the very pole of plannedness: oral art. The heterogeneity of data types in large-scale databanks notwithstanding (see above), the coverage of speech events along the spontaneity cline seems to have a certain tendency to stop short of oral art, presumably because it may depart in content and in form in ways that make it stand apart, in more significant ways, from other genres. There are important reasons for this phenomenon. For example, only few, if any, (often elderly) members of a speech community may be able to collaborate in fieldwork on oral art, as the necessary knowledge may be receding or is nearly lost. It is thus not uncommon that oral art traditions, even where (at least, marginally) surviving, are left out altogether from a language documentation project. (This situation is quite frequent in the Himalayan area.)

16. Note that this bias is more one of ideals and aims, rather than of quantitative predominance. Especially monologic speech, often much easier to transcribe, interlinearize and translate, remains a hallmark of many documentation corpora.

The choice to, at least in early stages, disregard oral traditions in a documentation project may also have to do with ideas around the longevity of such traditions. After all, the catholic mass in the Vatican is still held in Latin, and the Vedic tradition has survived for over 3000 years. These, however, are cases where societies have not undergone fundamentally disruptive change with regard to their cultural-spiritual frameworks. When such disruptions do occur, oral traditions are likely to be lost (long) before the varieties of everyday communication. In addition to the benefits for both historical linguistics and linguistic typology, then, I would like to argue for an inclusion of more conservative lects and genres, for the research-related reasons I have laid out in this article, but also from a broadly decolonial perspective, where research interests are adjusted to align, to the greatest extent possible, with a community's diverse interests and priorities (Leonard 2018, Tsikewa 2021).

Acknowledgements

Many thanks to Albert Tallai, Lou Guillaume, and Saida Asvarova for help with compiling the data in Tables 1, 2 and 6. I am very grateful for comments by Alec Coupe, Birgit Hellwig, Laura Becker, Naomi Peck and an anonymous reviewer, which helped streamline and improve this paper. All remaining errors are mine.

In more general terms, I want to thank Nikolaus Himmelmann. Nikolaus, it was sometime in 2007 that I heard you, for the first time, speak in class, back then at the Westfälische Wilhelms-Universität Münster. What you had to say captivated my attention in the fullest possible way. I am still here listening. Thank you for your friendship and for your mentorship. They mean more to me, and have guided me more, than I can say. And thank you, too, for listening.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following glosses are used:

CL	size-and-shape classifier
CONT	continuative aspect
CONV	converb
NOMLZ	nominalizer
POL	politeness marker.

References

Alpher, Barry. 1976. "Some Linguistic Innovations in Cape York and their Structural Correlates." In *Languages of Cape York*, ed. by Peter Sutton, 84–101. Canberra: AIAS.

Auer, Peter. 2018. "Dialect Change in Europe — Leveling and Convergence." In *The Handbook of Dialectology*, ed. by Charles Boberg, John Nerbonne, and Dominic Watt, 159–176. Wiley.

Becker, Laura, Matías Guzmán Naranjo, and Samira Ochs. 2023. "Socio-Linguistic Effects on Conditional Constructions: A Quantitative Typological Study." In *Sociolinguistic and Typological Perspectives on Language Variation*, ed. by Silvia Ballarè, and Guglielmo Inglese, 121–154. Berlin: De Gruyter.

Bisang, Walter. 2009. "On the Evolution of Complexity — sometimes less is more in East and mainland Southeast Asia." In *Language Complexity as an Evolving Variable*, ed. by Geoffrey Sampson, David Gil, and Peter Trudgill, 34–49. Oxford: Oxford University Press.

Blevins, Juliette. 2001. "Where have all the Onsets gone? Initial Consonant Loss in Australian Aboriginal Languages." In *Forty Years on: Ken Hale and Australian languages*, ed. by Jane Simpson, David Nash, Mary Laughren, Peter Austin, and Barry Alpher, 481–492. Canberra: The Australian National University.

Blevins, Juliette. 2007. "Endangered Sound Patterns: Three Perspectives on Theory and Description." *Language Documentation and Conservation* 1 (1): 1–16.

Bowern, Claire, John C. Smith, Betsy Sneller, Meredith Tamminga, Jadranka Gvozdanović, John Goldsmith, Götz Keydana, and Juliette Blevins. 2024. "Diachrony and Diachronica: 40@40." *Diachronica* 41(1). 127–140.

Bowern, Claire, Alan C.L. Yu, Salikoko S. Mufwene, Marlyse Baptista, Justin M. Power, Richard P. Meier, Bridget Drinka, Uta Reinöhl, and Simon Greenhill. 2025. "Diachrony and Diachronica: 40@40." *Diachronica* 42 (2): 137–160.

Bybee, Joan, and Östen Dahl. 1989. "The Creation of Tense and Aspect Systems in the Languages of the World." *Studies in Language* 13 (1): 51–103.

Chafe, Wallace. 1994. *Discourse, Consciousness and Time*. Chicago: The University of Chicago Press.

Culhane, Kirsten, Naomi Peck, and Uta Reinöhl. 2025. "No Evidence for Lexical stress in Kera'a, an Endangered Trans-Himalayan language." Proceedings of the Third International Conference on Tone and Intonation (TAI), 109–113.

DeLancey, Scott. 2014. "Sociolinguistic Typology in North East India: A Tale of Two Branches." *Journal of South Asian Languages and Linguistics* 1 (1): 59–82.

Dryer, Matthew & Martin Haspelmath (eds.) 2013. *WALS Online (v2020.4)*. Zenodo. (Available online at https://wals.info, Accessed on 2025-11-09.)

Evans, Nicholas. 2003. *Bininj Gun-Wok: A Pan-Dialectal Grammar of Mayali, Kunwinjku and Kune*. Canberra: Research School of Pacific and Asian Studies, The Australian National University.

Evans, Nick, and Stephen C. Levinson. 2009. "The Myth of Language Universals: Language Diversity and its Importance for Cognitive Science." *Behavioural and Brain Sciences* 32 (5): 429–448.

Fougeron, Cécile, and Patricia A. Keating. 1997. "Articulatory Strengthening at Edges of Prosodic Domains." *The Journal of the Acoustical Society of America* 101 (6): 3728–3740.

Fox, James (ed.). 1988. *To Speak in Pairs. Essays on the Ritual Languages of Eastern Indonesia.* Cambridge: Cambridge University Press.

Gaenszle, Martin. 2018. "Introduction." In *Ritual Speech in the Himalayas. Oral Texts and their Contexts*, ed. by Martin Gaenszle, 3–16. Cambridge, MA: Harvard University Press.

Hill, Clair E. B. 2018. Person Reference and Interaction in Umpila/Kuuku Ya'u Narrative. PhD Thesis, KU Leuven.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36: 161–195.

Hock, Hans Henrich. 1997. "Chronology or Genre? Problems in Vedic syntax." In *Inside the Texts — Beyond the Texts: New Approaches to the Study of the Vedas*, ed. by Michael Witzel, 103–126. Cambridge, MA: Harvard University Press.

Höder, Steffen. 2010. *Sprachausbau im Sprachkontakt. Syntaktischer Wandel im Altschwedischen*. Heidelberg: Winter.

Hopper, Paul. 1987. "Emergent Grammar." *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, 139–157.

Hopper, Paul J. 1991. "On Some Principles of Grammaticalization." In *Approaches to Grammaticalization*, vol. 1, ed. by Elizabeth C. Traugott, and Bernd Heine, 17–35. Amsterdam: John Benjamins.

Hualde, Jose Ignacio, and Jon Ortiz de Urbina. 2003. *A Grammar of Basque*. Berlin: Mouton de Gruyter.

Jakobson, Roman. 1966. "Grammatical Parallelism and its Russian Facet." *Language* 42: 398–429.

Keating, Patricia A., Taehong Cho, Cécile Fougeron, and Chai-Shune Hsu. 2003. "Domain-Initial Articulatory Strengthening in Four Languages." In *Phonetic Interpretation: Papers in Laboratory Phonology VI*, ed. by John Local, Richard Ogden, and Rosalind Temple, 145–163. Cambridge: Cambridge University Press.

Kelly, Lynne. 2015. "Primary Orality and Oral Mnemonic Technologies." In *Knowledge and Power in Prehistoric Societies: Orality, Memory and the Transmission of Culture*, ed. by Lynne Kelly, 36–61. Cambridge: Cambridge University Press.

Koch, Peter, and Wulf Oesterreicher. 1985. "Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte." *Romanistisches Jahrbuch* 36: 15–43.

Kölligan, Daniel, Claes Neuefeind, Uta Reinöhl, Patrick Sahle, Antje Casaretto, Anna Fischer, Börge Kiss, Natalie Korobzow, Jürgen Rolshoven, Jakob Halfmann, and Francisco Mondaca. Ongoing. "VedaWeb. Online Research Platform for Old Indic Texts." University of Cologne. https://vedaweb.uni-koeln.de, accessed 09/10/2025.

Kuznetsova, Natalia, Cormac Anderson, and Shelece Easterday (eds.). 2025. *Rarities in Phonetics and Phonology: Structural, Typological, Evolutionary, and Social dimensions*. Berlin: Language Science Press.

Lehmann, Nico, Vahid Mortezapour, Jozina Vander Klok, Zahra Farokhnejad, David Müller, Elisabeth Verhoeven, and Aria Adli. 2025. "Lang*Reg corpus: Documenting Intra-Speaker Variation Across Languages and Registers." *Language Documentation and Conservation* 19: 40–66. https://hdl.handle.net/10125/74810

Leonard, Wesley Y. 2018. "Reflections on (De)colonialism in Language documentation." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 55–65. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24808

Levinson, Stephen C. 2006. "On the Human 'Interaction Engine'." In *Roots of Human Sociality: Culture, Cognition and Interaction*, ed. by Nick J. Enfield, and Stephen C. Levinson, 39–69. Oxford: Berg.

Lupyan, Gary, and Rick Dale. 2010. "Language Structure Is Partly Determined by Social Structure." *PLoS ONE* 5(1): e8559.

Maas, Utz. 2010. "Literat und orat: Grundbegriffe der Analyse geschriebener und gesprochener Sprache." *Grazer Linguistische Studien* 73: 21–150.

Masica, Colin. 1991. *The Indo-Aryan Languages*. Cambridge: Cambridge University Press.

Meakins, Felicity, Xia Hua, Cassandra Algy, and Lindell Bromham. 2019. "Birth of a Contact Language did not Favor Simplification." *Language* 95(2): 294–332.

Meyerhoff, Miriam. 2019. "Unnatural Bedfellows? The Sociolinguistic Analysis of Variation and Language Documentation." *Journal of the Royal Society of New Zealand*, 49 (2): 229–241.

Moran, Steven, Nicholas A. Lester, and Eitan Grossman. 2021. "Inferring Recent Evolutionary Changes in Speech Sounds." *Philosophical Transaction of the Royal Society B: Biological Sciences* 376: 20200198.

Ochs, Eleanor. 1979. "Planned and Unplanned Discourse." In *Discourse and Syntax*, ed. by Talmy Givon, 51–80. New York: Academic Press.

Ong, Walter. 1982. *Orality and Literacy: The Technologizing of the Word*. London and New York: Routledge.

Reinöhl, Uta. 2016. *Grammaticalization and the Rise of Configurationality in Indo-Aryan*. Oxford: Oxford University Press.

Reinöhl, Uta. 2022. "Locating Kera'a (Idu Mishmi) in Its Linguistic Neighbourhood: Evidence from Dialectology." In *Ethnolinguistic Prehistory of the Eastern Himalaya*, ed. by Mark W. Post, Stephen Morey, and Toni Huber 232–263. Leiden: Brill.

Reinöhl, Uta. 2024. *Documentation of Igu*. Data Center for the Humanities.

Reinöhl, Uta. To appear. "Poetic Structures and Strategies of Igu — The Shamanic Language of the Kera'a." *Oral Tradition*.

Reinöhl, Uta. In prep. "The historical lab approach to language change."

Reinöhl, Uta, and Antje Casaretto. 2018. When Grammaticalization does NOT occur. Prosody-Syntax Mismatches in Indo-Aryan. *Diachronica* 35 (2): 238–276.

Reinöhl, Uta, Daniel Kölligan, Börge Kiss, Francisco Mondaca, Claes Neuefeind, and Patrick Sahle. 2018. "VedaWeb — eine webbasierte Plattform für die Erforschung altindischer Texte." In *Book of Abstracts der 5. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd)*, Köln, 26.2–2.3.2018, 485–486.

Reinöhl, Uta, Pachu Pulu, and Usha Wallner. 2025. "A Sketch Grammar of Igu, the Shamanic Language of the Kera'a." *Himalayan Linguistics* 24 (1): 58–87.

Reinöhl, Uta, Kirsten Culhane, and Naomi Peck. Under review. "The mystery of word-initial consonant loss in Kera'a."

Reinöhl, Uta, and Naomi Peck. In prep. "A Sketch Grammar of Kera'a."

Skirgård, Hedvig, Hannah J. Haynie, Harald Hammarström et al. 2023. *Grambank v1.0 (v1.0)*. Zenodo.

Stivers, Tanya, Nick Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. "Universals and Cultural Variation in Turn-Taking in Conversation." *Proceedings of the National Academy of Sciences. U.S.A.* 106 (26): 10587–10592,

Trudgill, Peter. 2008. "Vernacular Universals and the Sociolinguistic Typology of English Dialects." In *Vernacular Universals and Language Contacts*, ed. by Markku Filppula, Juhani Klemola, Heli Paulasto, 304–322. London and New York: Routledge.

Trudgill, Peter. 2011. *Sociolinguistic Typology. Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Tsikewa, Adrienne. 2021. "Reimagining the Current Praxis of Field Linguistics Training: Decolonial Considerations." *Language* 97 (4): e293–e319.

CHAPTER 13

Songs in Eipo, Yale, Awiakay and Meakambut, four Papuan languages

Volker Heeschen & Darja Hoenigman
Ludwig-Maximilians-Universität München | The Australian National University

Songs are expressions of creativity and emotion that have great intrinsic value not only to the cultures that produced them, but also to academics from fields such as linguistics and anthropology, who seek to gain deep and meaningful insights into these cultures. Here, we analyse some fascinating aspects of the language of song and related speech acts in four languages of New Guinea. Both structural and semantic aspects of song language are discussed, in order to show the differences between it and everyday speech. Repetitions, metaphor, allusion and unintelligibility are deliberate and prominent features in song language, and this can pose difficult, but not insurmountable, challenges to outsiders who wish to understand (and by extension, translate) the text of the songs.

Keywords: multilingualism, metaphor, mockery, parallelism, translation, song language, unintelligibility

1. Introduction

Songs are ubiquitous across human cultures, and it has been widely observed that the language of song can often deviate greatly from the norms of spoken language. Song language from various parts of New Guinea has been found to share a number of distinct features, such as repetitions (of single words, names, phrases or entire lines), parallelism (repetition that includes an element of variability), modifications (phonological, lexico-grammatical, semantic or combinations thereof) to everyday language and unintelligibility (the intentional use of foreign or archaic material) (Rumsey & Niles, to appear). Beyond the structural aspects of the language, the meanings of songs also demand special attention. This is because:

https://doi.org/10.1075/slcs.240.13hee
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

> ...they frequently have distinct readings from that which would be assigned to them on the basis of straightforward compositional semantic principles, in the same way that phrasal idioms do. As well, song often conveys more meanings, significances and resonances than an ordinary speech utterance, and in this respect songs resemble poetry. In songs, meanings are encoded not only in the text, but also through a complex interaction of the music, dance, visual designs and their associations with specialized knowledge, including knowledge about the context of performance. (Turpin & Stebbins 2010: 2)

With respect to their status in language documentation, songs can be said to lie close to the 'planned' pole of the spontaneity spectrum, and to occupy the space between 'monologic' and 'ritual' speech genres (Himmelmann 1998). Songs and related speech acts are themselves highly diverse, and as will be seen in the following sections, include vocal accompaniment to truly ritual activities, creative compositions to evoke, or in response to, strong emotions, as well as conventionalised dialogic utterances intended to shame, complain or chide. Therefore, songs and related speech acts have a high intrinsic value not only as objects within a comprehensive language documentation program, but also as sources of insights that can help inform an ethnography of the culture in question (Franchetto 2006; Himmelmann 2006).

This chapter takes a closer look at the extraordinary (compared to everyday speech) features of the language used in songs and songlike utterances in four different languages spoken on the island of New Guinea. Two (Eipo and Yale) are spoken in the eastern highlands of the Indonesian province Papua Pegunungan, while the other two (Awiakay and Meakambut) are spoken in the East Sepik Province of Papua New Guinea. Due to space constraints, only some of the features mentioned above will be discussed in this paper; among the structural features, parallelism and unintelligibility are featured prominently, while the exploration of meaning creation focuses on allusion and metaphor on the one hand, and subtle/indirect criticism and mockery through song on the other. Two first person voices are used in the following text, which reflect the mixed authorship of the chapter. Simply put, the first person narrator in the Eipo/Yale sections below represent the observations of Volker Heeschen, while the 'I' of the Awiakay/Meakambut sections is Darja Hoenigman. Apart from the evident academic aspects of the following discussion, this contribution could also be regarded as a means of conjuring up memories of some of the first moments of fieldwork, when the fieldworker begins to think about his/her work as being meaningful and joyful.

2. Setting the scene

2.1 The Mek languages

Eipo and Yale are members of the Mek language family, whose homelands lie in the centre of the island of New Guinea, on the Indonesian side of the international border (hence in the province of Papua Pegunungan). They were formerly known as the Goliath languages, and since 1978 have been called the Mek languages (Godschalk 1993; Schiefenhövel 1976). The Eipo live in the centre of the Mek area, the Yalenang are the westernmost members of the Mek tribes (cf. Heeschen 1998: 17). The Mek languages form a family of closely related languages within the Trans New Guinea Phylum. The percentage of shared vocabulary never falls below 60%. The main group of languages (or dialect clusters) are: (1) In the north-west, the Yale language and neighbouring dialects, (2) in the south-west, the dialects spoken around Korapun, Sela, and, probably Nalca, (3) in the centre, Eipo, Una, Tanime, and Langda, (4) in the east all dialects spoken between Bime and Okbap. Whether the Momuna language spoken somewhere south of Wamena forms part of the Mek family remains to be determined.

Tales of origin, settlement patterns as well as dialect geography point to an area in the south-east of the Mek region from where the people spread, migrated and colonized the large valleys. One may assume that speakers of Mek languages and dialects maintained social and cultural contacts with speakers of the Ok languages in the east and speakers of the Dani languages in the west. In a not too-distant past, Mek speakers occupied a territory beyond the present boundaries of their area.

Permanent contact with the outside world was established in 1974. People made peace around 1976, the stone-adzes disappeared, traditional life stopped all at once, and in 1981 about the first third of the population was baptized. In 2014 the translation of the New Testament was finished and the people of the hamlet Sedekahi were baptized due to the missionary work of Mike Martin. Martin, who continued his predecessor Steinbring's work, started around 2002, and mainly lived in Sentani, a city at the coast.

2.2 Awiakay and Meakambut land and language

Awiakay land covers approximately 300 km² of lowland rainforest and freshwater swamp forest area around the upper Konmei River in the southern flood plain of the Sepik River in the East Sepik Province of Papua New Guinea. Its southern part stretches into the northern Fringe Highlands of the New Guinea Central Range, and touches the Maramuni River to the south-east. Kanjimei village, where the

Awiakay are settled, is located in the north of this territory. The entire Meakambut community numbers only 64 individuals, and they are neighbours of the Awiakay. They are not sedentary like their neighbours, but constantly move between camps and rock-shelters in the vicinity of sago stands, following mature sago palms and foraging in their mountainous limestone territory.

The Awiakay have been nominal Catholics since the 1950, but have built their first church in 2004. It resembles the churches in neighbouring villages and is built on the ground. There are two groves of rubber trees at the edge of the village, which resulted from extension work by a 'district agricultural officer' who in the 1970s introduced the growing of rubber. The easternmost part of the village is a cleared area used as a soccer field.

The Awiakay have often moved their settlement and today's Kanjimei was only established some sixty years ago when people moved to the north in order to protect the borders of their land from the neighbouring Ambonwari and Imanmeri.[1] Until the early 1960s people lived in Marinyam, about 10km upriver from Kanjimei. The former village is nowadays deserted and overgrown with grass and trees. It serves as a bush camp and there are a few houses there where people can stay overnight when they are on their way to or from the neighbouring village of Asangamut, or when they go hunting. Previous settlements of the Awiakay are mentioned in their Myth of Origin (Hoenigman 2007: 242).

Kanjimei nowadays consists of seventy-six family houses, one 'single-man's house' and my house, which stands in the middle of the village.

Awiakay and Meakambut are Papuan languages, both belonging to the small Arafundi family. Awiakay is spoken by about 500 people living in Kanjimei village in the East Sepik Province of Papua New Guinea. Meakambut has 64 speakers, living in the territory south of Awiakay land. In Foley (1986) originally, the Arafundi languages were described as a single language with a high degree of dialect variation. However, my investigation of lexical similarities and mutual intelligibility indicate a degree of diversity that points to a distinction between seven or eight different languages (Hoenigman to appear a).

1. It should be noted that the first location of Kanjimei was slightly more towards the northwest, down the Konmei, and that the present location was finally chosen after some disputes with the Ambonwari. Occasionally moving the location of their settlements seems to be rather typical of all people in the Sepik Hills area (cf. Telban 1998: 15; Toyoda 1998: 4).

3. Personal journeys and research philosophies

3.1 Heeschen

In Heeschen et al. (2015) I described my first interest in the peoples and languages of the South Pacific, the duration of the interests while being a scholar and a student, the contact with the founders of our project in New Guinea as well as our knowledge of what was first called the Goliath people and their languages, later the Mek people and the Mek languages. The flight from Eipomek to Pagai and trekking in the area south of Pagai was just one step in our activities, which ended in July 2014, aiming at getting to know almost all areas of the Mek area, as well as documenting the diversity of the Mek languages (see Heeschen et al. 2015: 7–19).

Linguistic work on the Yale language started in May 1979 while that on the Eipo could be continued either by having Eipo guests in Kosarek (where Yale is spoken) or visiting Eipomek (where Eipo is spoken). I stayed in Kosarek until November 1981, revisiting Kosarek as well as Eipomek 17 times, between 1982 and 2014, for periods of up to four months; thus I spent more than five years in the Mek language area.

From the beginning my interests were not limited to pure linguistics; however, grammar, dictionary and texts were not neglected. I investigated a wide range of topics during fieldwork, namely typological questions related to, and arising from, observing characteristics of ways of speaking, the limited size of clauses qua speech units, as well as the gift of speakers to constantly formulate corrections, to improve wordings or to additionally explain events and circumstances. Besides stepping into fields other than those recommended for linguists I scrutinized myths, migration stories, accounts of war, narrations of settlement and everyday life, contents and meaning of song and dancing song texts, and also interviewed native speakers about cultural values and social rules of behaviour.

3.2 Hoenigman

Before I arrived in Kanjimei in 2004, no research had been done in the village and linguists were unaware of the existence of the Awiakay language. Wurm and Hattori's Atlas of Languages of the Pacific (1981) shows the location of Kanjimei as an uninhabited area (though close to Alfendio, i.e. Arafundi language). Even in the 2004 SIL East Sepik Province Language Map, the northern border of Arafundi does not reach Kanjimei. In Laycock's (1973) survey of the languages of the Sepik region, Kanjimei village is not mentioned at all. When I first arrived there it was believed that the language spoken in the village was either Miyak (a Yuat River language) or the language of Imanmeri village (an Arafundi language), both of which are neighbouring languages of Awiakay. The first documentation of

Awiakay is in my unpublished MA thesis 'Language and Myth in Kanjimei, East Sepik Province, Papua New Guinea' (Hoenigman 2007).

During this first field trip, I was intrigued by the unclear linguistic picture in the area, and I set off for a couple of trips on the Arafundi river. This allowed me to record basic wordlists for the different lects spoken in the villages of Imanmeri, Wamblamas, Yamandɨm, Awim, Imboin (including Karɨmba), and the seminomadic Meakambut. Starting in 2004, I have conducted 14 fieldtrips to the Arafundi area, where I have spent a total of 49 months.

Since 2015, I have also been visiting the neighbouring Meakambut, whom I first contacted with the help of some Awiakay speakers. Their Arafundi and Karawari neighbours share the same origin myth according to which the Meakambut are the guardians of the place of origin, namely a stone 'vagina' (limestone cave) in Mt. Kopao.

Participant observation has been a key methodology throughout my fieldwork. I would attempt to be present, often with a video camera, for a wide range of interpersonal and/or public interactions, including conversations, quarrels, religious ceremonies, hunting and foraging trips, and so on. My presence, and the presence of the video camera was welcomed, and even encouraged, by both Awiakay and Meakambut. With the cooperation of the people of Kanjimei, I was able to conduct an ecologically (*sensu* Haugen, in Mühlhäusler 1996) sensitive documentation of various Awiakay speech registers. In line with Garner's (2005) understanding of ecological philosophy, my approach was holistic (concerned with complexity, diversity and interrelationships), dynamic (concerned with the unpredictability or creativity inherent in language), interactive (concerned with verbal and non-verbal aspects of communication, as well as the speakers' interactions with the community as a whole) and situated (concerned with the context — including the physical environment — in which language is used).

4. Songs in Eipo, Yale, Awiakay, and Meakambut — overview

4.1 Eipo and Yale songs — overview

The Eipo sing *dite*, 'songs', the Yale *kelem*, 'courting songs'. The *dite* are sung alone, in very small groups, for instance when two or three women sit together repairing netbags, when a man carves arrow tips, or when two boys sit together and look around. The *kelem* are said to be sung by young men after dusk or during the night, when people sit in the family huts or the men in the men's house. The audience may be the whole hamlet; as the relation to the Angguruk's people's songs is not clear, and as most of them are sung in the Angguruk language (a member of

the Dani family, which I do not know and which most of the Yale people do not understand), I will not discuss them here. However, the reader can consult the section on Awiakay song language (below) for a more thorough treatment of this phenomenon.

Asked what a *dit* 'song' is, the Eipo say that whenever a woman has had a love affair, she will compose a song. Each song should have an individual author as well as a distinct text and melody. However, besides love songs there are other songs composed by male authors who criticize lazy lovers, ridicule those who do not give and share, express sorrow and longing, or just derive pleasure from looking at ornaments or observe how leaves hang down in the rain. The *kwemdina dite* 'creation songs', also referred to as *Yaltapenang dite* 'the songs of the Yaltape people', are nothing but the 'pop songs' of the older generation. These songs mention the paths across the mountains to the neighbouring valleys, encounters along the way, trading-partnerships, and, as the dead are said to gather at the high places, they also serve as mourning-songs.

The Eipo term *mot* and the Yale word *sebna* signify the whole of a dancing event, the act of dancing, and the text of a dancing-song. The Yale word *mod*, corresponding to Eipo *mot*, means special 'sacred places' as well as 'polished stones', which were kept in the men's houses, and inserted into the ground to assure the consolidation of the earth. Indeed, the older dancing-song types mention sacred places and mountains, so that the relation between dancing and worshiping the places of origin and the appearance of cultural items is suggested. In particular, the songs sung during the initiation ceremonies connect places with worship, for instance by identifying waterfalls as hiding places of ancestors, pools as places of origins of pigs. The *nassa* headbands, the cassowary and some cuscus species as totems are also celebrated.

The Eipo and Yalenang mainly dance on two occasions: during exchange or pig festivals and after triumphantly returning from warfare. The importance of other events like the handing over of a compensation gift or goods from the distant lowlands or the final exchange for establishing a marriage can be stressed by dancing. Starting from one's own hamlet, singing and dancing along one's way to the hosts, as well as decorating oneself, is experienced as a *fait social total*, the setting off from dirt and everyday routines to worship, beauty, riches, and renewal of life.

In these songs, the verbs are often stripped of their suffixes referring to person, tense, and mood. All movements, objects, places, rivers, birds, plants, and phenomena of nature are laden with, and symbolic of, social meanings as well as veiled references to protagonists, values and emotions. A place name can stand for clans who own the place, someone with his hands behind his back or who moves backward can imply a man who is idle or does not pay the bride wealth, while a bridge can represent encounters and the meeting-point of lovers.

4.2 Awiakay and Meakambut songs — overview

The Awiakay state emphatically that *Kaunjambi*, an all-night song/dance cycle, is iconically Awiakay. All-night dances are also performed by some neighbouring groups, such as Ambonwari, Imanmeri and Awim — but not Asangamut —, but according to the Awiakay, *Kaunjambi* is unique, and belongs to their community alone. Of the song/dance cycles mentioned above, the *yamɨn siria* performed in Ambonwari seems to be the most similar to *Kaunjambi* (Telban 1998: 196–221). However, the *Kaunjambi* songs have a reduced narrative component, and contain more untranslatable elements than those of *yamɨn siria*. I elaborate on this point in Section 6.2. While women do not actively sing *Kaunjambi* songs, the performance does involve the whole village. *Kaunjambi* is performed either in honour of an individual (e.g., a special guest) or a local person's achievement (e.g., completing a teacher's course), or as a celebration of something important to the whole community, such as reconciliation, or, in more recent times, a new school, or a group of children receiving the Christian sacraments of First Confession and Communion.

Kaunjambi is one of the three Awiakay all-night song/dance cycles, the other two being *supim* and *telek endinga*. While *supim* used to be sung in the men's house during initiation rites, *telek endinga* 'war song' was performed after headhunting raids (Hoenigman 2007: 34, 53). Along with their associated rituals, both dances/song-cycles were banned by the early visiting missionaries, and Awiakay people believed that if they performed them again, the spirits would cripple the legs of the uninitiated children.

Although women do not sing *Kaunjambi*, as mentioned above, they are actually very good song composers — and singers — of a song genre called *nam endiŋga* 'women's songs', which is melodically and rhythmically closely related to *Kaunjambi*. Women compose songs in order to seduce men who are invariably so touched by them that many of these attempts actually do end up in a marriage. Some of the women's songs are remembered and sung for generations, and everybody remembers who composed them and for whom.

It took me a long time before I heard a Meakambut song, and it was only later that I realised that the Meakambut do not usually compose songs in their own language. The stated reason: *Em bai no kamap gut, ia!* 'It won't sound good!' While they hardly ever code-switch between the local lingua franca Tok Pisin (TP) and Meakambut, and use very few TP loanwords, I would often hear the Meakambut singing in TP (mostly the popular TP songs that they learn from SD cards). And whenever I heard someone singing in another language, and asked if that was a Meakambut song, they would reply that it was in neighbouring languages such as Maramuni (an Engan dialect) or Anday (another Arafundi lan-

guage). It was only well into my fieldwork that I was given a clear explanation: "When [people] would like to sing ancestral songs, they'd compose them in the language of Andambɨt [Anday], and the Andambɨt would compose their songs in our language. One can't compose songs in one's own language. It's custom. Our ancestors did it this way, and this is how it is. If you composed [a song] in your own language, it just wouldn't sound good."

5. Parallelism in verbal art

Verbal art in many cultures around the world makes use of the device known as *parallelismus membrorum*, or simply, parallelism. This phenomenon commonly occurs in the languages discussed here. In textual (as distinct from musical or gestural) parallelism, one or more linguistic elements within a line (words, morphemes, phonemes, syntactic constructions) are repeated in another line, while one or more other elements (A1, (A2, ...)) is/are replaced with other, non-repeating element(s) (B1, (B2...)). If the line is repeated again, element B(1) is replaced with C(1), and so on. Hence Rumsey's definition of parallelism as "the ordered interplay of repetition and variation" (Rumsey 2007: 261). A simple example involving only one variable element is:

d e f g h A
d e f g h B
d e f g h C

A, B and C stand out from the repeated elements in each line. They are distinct, unique elements, but also share a special relationship, as they appear at the same location with respect to the repeated elements. This relationship was termed 'equivalence' by Jakobson (1966: 400), after Lowth (1779: x–xi). There is always some association between these variables in a parallel set, which means that they are perceived to be closely connected, although not being the same ('salt' and 'pepper' in the Western mind for example). Nor are they exact opposites (Hoenigman 2015a).

The associations between variables in a parallel set reflect cultural conceptualizations, which are shared by members of a given community. A simple example from the Awiakay *Kaunjambi* is given in (1):

(1) Awiakay
wi *mendus poko-y*
quicklime tongue hit-FUT.3PL
'quicklime is burning the tongue'

komboŋa mendus poko-y
shell tongue hit-FUT.3PL
'the shell is burning the tongue'

The repeated words in this case are *mendus pokoy*, and the non-repeated ones — the variables — are *wi* 'quicklime' and *komboŋa* 'shell'. While the association between quicklime and shell may not be immediately apparent to an outsider, it is obvious to the Awiakay, and to all those New Guineans who regularly produce lime for chewing with betel nut by burning the freshwater mussel shells which they collect from the river. Note that the set of variables does not need to be binary; indeed, 73% of all parallelisms in *Kaunjambi* feature parallel sets larger than couplets, with the largest comprising eight parallel lines, with eight different variables.

The same verbal device can be seen in the following excerpt from a Yale initiation song in (2), as well as in the Eipo songs presented in Section 6.1:

(2) Yale
Ya'ol domdom-o,
'at the Ya'ol river singing the chorus, oh,'
Kweleab domdom-o
'at the Kweleab river singing the chorus, oh!'

The above excerpt mentions two river names, and this happens to be a very common occurrence in Awiakay songs as well. About one third of all *Kaunjambi* songs include parallel sets in which the variables are place names. Names of creeks, sometimes interchanged with hills, and names of pools, which are common spirit places, commonly appear as sets of parallel terms.

6. Song language in Eipo, Yale, Awiakay, and Meakambut

6.1 Symbolism, evocation and allegory in Mek and Yale songs

The day we arrived in Eipomek we heard the Eipo sing a song. Wherever an Eipo sits and works he or she may start singing. Whenever we met an Eipo while travelling, he or she might emphatically greet us, the greeting turning into a song. A greeting formula like *nani Fatere* 'my father Fatere (i.e., Volker)' may give the impression of starting a song. (3) is one of the first greeting songs encountered. In this and the following examples, the name of the song appear(s) first in the heading, followed by the name(s) of the author(s) (if known).

(3) Eipo, *Malyonang kilon dit* (Greeting song of the Malyo clan), sung by Buk and Bingde

Nanin waue, na kaluse,
my.mother waue.beetle my kaluse.beetle
'my mother *waue* beetle, my *kaluse* beetle,'

Kacang kolungdine, Sungkon belingdine,
Kacang descending Sungkon descending
'coming down from Kacang river, descended from Sungkon hamlet,'

Nanin kole, neka waue,
my.mother kole my.friend waue.beetle
'my mother *kole* snake, my friend *waue* beetle,'

Kacang koukwe, Sungkon koukwe
Kacang beauty Sungkon beauty
'oh beauty from Kacang river! oh beauty from Sungkon hamlet!'

The beetles, the river and the hamlet are classified as belonging to the Malyo clan and to the place of the Malyo people's origin. They fit into what is property or a characteristic feature of the Malyo clan.

Just as I discovered the Eipo greeting songs by chance, so were the Yaltapenang songs: I made a recording of the songs sung by two boys sitting in front of my hut, and the songs turned out to be almost the complete repertoire of the Yaltapenang songs. The Yaltape people are the ancestors of the Mekdoman clan of Dingerkon. The songs describe the migrations, the wanderings and life in the forest. The songs could be interpreted as containing veiled wording for love affairs, they could be regarded as expressions of mourning because the high forests are the retreat of the spirits, or they could convey the joy a man feels when he returns home through rain and over steep slopes. In the following, the author Bowungde expresses a sad mood.

(4) Eipo, *Yaltapenang dit* (Yaltapenang song), sung by Bowungde

Keting kanukde, dongobre,
sun it.curls.up it.lies.on
'The sun has curled up (the leaves), it shines on them,'

Ure fu-nang, eidam, Lire fu-nang eidam,
Ure headwaters above Lire headwaters above
'the people from the Ure river up there, the people from the headwaters of the Lire up there,'

doba-bum barebre, mo-bum donokde, amea!
cloud-within he.is.covered rain-within he.is.thrown friend
'he is covered by clouds, he is exposed to the rain, friend!'

Keting donokde, keting marabre, i-dide-yo!
sun he.is.thrown sun he.is.wounded up-look-yo!
'He is exposed to the sun, the sun shines on him, look up!'

Keting kokde, keting tekde.
sun he.is.shrunk sun he.is.hidden
'He shrinks in the sunlight, he stands in the shadow.'

Sumsum keta dinmarmar, yane subumal yanibne.
Sumsum flower he.is.eating foot I.have.broken.through
'He is eating the flowers of the *sumsum* bush, my foot has broken through the rotten wood of the path.'

Bibulkau yukdobre, kambulkau kwanikdobre.
Bibulkau it.hangs.down kambulkau it.breaks.off.
'The (leaves of the) *bibulkau* bush hang down, the (leaves of the) *kambulkau* bush shed.'

The dead are exposed to sun and rain and they come together with the ancestors in the headwaters of the rivers. They turn into birds which eat the flowers and fruits of the plants. Sadness is expressed through the 'rotten wood of the path' and the leaves which hang down. The leaves of the plants which are mentioned in the song are known for wilting quickly in the sun.

Love songs are also full of allusions, as in the following *Kolubkolub dit* in (5) sung by Oleto:

(5) Eipo, *Kolubkolub dit* (Love song), sung by Oleto

Kokiye dam, kokbare dam,
PLANT at PLANT at
'By the *kokiye* fern, by the *kobade* fern,'

na lebnuke, na nuknuke.
me they.said.me me they.assigned.me
'they said to me, they assigned him to me.'

An-ning guma, na-ning anya
your-own not my-own this
'"It is not yours, it is mine"'

na lebnuke, na nuknuke.
me they.said.me me they.assigned.me
'they said to me, they assigned him to me.'

Kidiknamne, banabnamne.
I.will.be.lost I.will.go
'I will lose my way, I will go,'

Na dukiye, na kokaye
my dukiye my kokaye
'my *dukiye*, my *kokbaye*,'

na lebnuke, na nuknuke.
me they.said.me me they.assigned.me
'they said to me, they assigned him to me.'

Se kayape, nay kayape
woe friends oh friends
'Woe, friends! Oh, friends!'

Kidiknamne, banabnamne.
I.will.be.lost I.will.go
'I will lose my way, I will go.'

Se, fotong-o, nay atiye
woe hair oh hamlet
'Woe, the hair(y fern)! Oh, the hamlet!'

na lebnuke, na nuknuke.
me they.said.me me they.assigned.me
'They said to me, they assigned him to me.'

Table ya metebume,
PLANT axe you.cut.up
'You have cut up the *table* tree with the axe,'

Like ya kwinebume,
PLANT axe you.chopped.up
'you have chopped up the *like* tree with an axe.'

Koubuknamne, webuknamne,
I.will.plant, I.will.make.garden
'I will plant, I will make gardens,'

Lung aik dam, fan aik dam.
PLANT hut at wild.one hut at
'at the *Lung*-grass hut, at the hut of the wild one.'

Kidiknamne, banabnamne.
I.will.be.lost I.will.go
'I will lose my way, I will go.'

Ya metebum, ya kotobum.
axe you.cut.up axe you.pounded
'You have chopped with the axe, you have pounded with the axe.'

Urye ya metebum wilalyam!
PLANT axe you.cut.up hit
'Cutting up the *urye* tree, hit it with the axe!'

Dibrenangkin, kwinibnangkin, delebnangkin.
I.will.see.you I.will.bend.you I.will.pull.you
'I will see you, I will bend you, I will pull you.'

Fin dam sekne, nab dam sekne.
reed at tinder PLANT at tinder
'The tinder of the reed, the tinder of the *nab*-plant.'

Cebroblyame, sekdoblyame!
break ignite
'Break it and ignite it!'

Ape Lunge, ape fane
dear PLANT dear wild.one
'Dear *Lunge* grass, dear wild one,'

Kwinibnangkin, delebnangkin.
I.will.bend.you I.will.pull.you
'I will bend you, I will pull you.'

Kidikse se wicape, nay kayape!
being.lost woe brothers oh friends
'Oh I am lost, relatives, oh friends!'

Kidiknamne, banabnamne.
I.will.be.lost I.will.go
'I will lose myself, I will go.'

Dibrenangkin, mebdeibnangkin.
I.will.see.you, I.will.plant.you
'I will see you, I will plant you.'

Se wicape, nay kayape.
woe brothers oh friends
'Oh relatives, oh friends.'

Fin dam sekne, nab dam sekne
reed at tinder PLANT at tinder
'The tinder of the reed, the tinder of the *nab*-plant.'

Toubuk dablyam, lekuke dablyam!
cut you.shall.burn split you.shall.burn
'Cut it and burn it, tear it and burn it!'

Subkor welen, korbik welen.
PLANT climbing PLANT climbing
'Climbing up the *subkor* tree, climbing up the *korbik* tree,'

ati teleb welebnamne, yanibnamne,
hamlet good I.will.climb I.will.step
'I will go up to a good hamlet, I will well step there,'

ape yane, ape fane.
dear coming dear wild.one
'sweet arrival, dear wild one.'

Lines 24–25 (shown in bold) provide the key for the contents: the author and singer Oleto had made a mistake, and a man she did not love was assigned to her. As is common in such songs, the trees and the fern denote a man; cf. line 10, where the hairiness of the fern is mentioned. But Oleto will make yet another mistake, and in the garden where she goes to perform various tasks, she will find a lover, a *fan* 'wild plant' or 'wild one' who was not assigned to her. The *furume si* 'the characterizing name' of the lover is subtly mentioned in line 10; this is *ati*, which in the Larye and Tani dialects means 'hamlet'. Thus, the meaning of lines 31–32 becomes clear: after losing her way she will go to another, better, hamlet, namely to her Ati. The lines containing *fane* indicate the name of the lover, as his name before initiation was Fane 'wild plant' (as opposed to a cultivated plant). The lines beginning with *Fin dam sekne* have an alternative meaning, namely 'exchanging or pulling off what is at the elbow', *i.e.* the armband. The implication is that Otelo wishes to love him intensely and remain in his company. Line 28 also evokes the harvesting of taro, and makes a veiled reference to the small and inefficient 'taro-penis' of the husband which Oleto would like to rip off and burn. These lines also imply that Oleto has had enough of the old garden plots, and so with the old marriage.

Dancing-songs

The questions of form, authorship and occasion were mentioned above. The following examples and comments are taken from former publications (Heeschen 1990, 2005). In the texts published and in texts being prepared to be published, dancing-songs and songs are pervasive, some myths explain former events, a fairy tale may end with a song or with the performance of a dancing-song. During initiations and ceremonies like the production of the *kelabi* shield, sung moments are singled out, and the initiation of the boys and dancing-songs culminate in the climax of sung events. The dancing-song in (6) occurs during the first initiation, when the boys are dressed with bundles of green *sabuda* leaves:

(6) Yale, *bed*-type dance

Ya'ol domdom-o, Kweleab domdom-o,
'At the Ya'ol river singing the chorus, oh, at the Kweleab river singing the chorus, oh,'

sabuda sae-a ikiok-di-e,
'as the *sabuda*-leaves scratched your hand, eh,'

dak sabuda-o, sale ikiksele,
'the *dak sabuda*-leaf, oh, it has scratched and touched you,'

Ya'ol domdom-o, Kweleab domdom-o,
'at the Ya'ol river singing the chorus, oh, at the Kweleab river singing the chorus, oh,'
a'un ba me-a ngikngika walelale-o!
'your boy, too, shall chuckle, oh!'

The leaves are a sign of 'scratching' off the dirt of the pre-initiation life. The boys shall start a new life by 'laughing' and 'singing' and being active. *Domdom* not only means 'sing in a chorus', but also 'kick up a noise'.

Ancestors are addressed in the dancing-songs. The Dabo-dance in (7), the lowlands dance, mentions Yalenye, 'the man of the east', the ancestor of the widespread Deal clan. Dawn, sun and light are welcomed. The rope (*tape*) goes from the top of the men's houses to the top of the mountains, along that rope the spirits shall go home.

(7) Eipo, *dabo*-dancing tune
Im bererimna tape aidam daknenibrum,
'The rope from the hut along the sky, untie it and roll it up,'
berentam tarwe dangkure, bingkube!
'when the morning comes, we will go!'

Images of nature are observed, interpreted as signs that are laden with symbolic meanings. The falling drops of rain from the trees after a cuscus or a bird move through the foliage evoke a sense of sadness. In the (Eipo) Yaltapenang songs some lines are *bisik ban yupe* 'talk about walking'. Talk about smoke indicates places where other people have rested, and hints that one has already gone a long way and needs a rest. Admonishments are common themes in the dancing-songs, which state that women shall not go to another hamlet, they shall not touch small objects, such as the tiny eggs of insects, because a landslide could be caused by (Eipo) *duke kurukna* 'egg loosening'. A scandal could be caused by men who approach a group of hard working women; here, what is a love-affair for one party, can be a source of mockery for the other. Multi-functionality of the texts is a characteristic feature of the dancing-songs, and the mention of something as apparently trivial as empty net bags could mean that the guests have eaten nothing or that the hosts are aware of the fact that they are obliged to give.

6.2 Language, metaphor and allusion in the Awiakay *Kaunjambi*

The Awiakay maintain that *Kaunjambi* is a song-cycle that has been composed by spirits in the distant past. They say this explains why certain parts of it are in an untranslatable *tasia menda* 'spirit language'. Other untranslatable parts include

traces of neighbouring languages such as the related Imanmeri language, the unrelated Karawari of the neighbouring Ambonwari, and in a couple of examples even the unrelated Miyak, a Yuat language spoken in Asangamut. Yet other parts of *Kaunjambi* are described as *nasepamgoy menda* 'the language of the past'; below, I refer to this as 'archaic Awiakay'. Borut Telban (pers. comm., Dec. 2012) who works with the neighbouring Karawari-speaking Ambonwari, was able to confirm that some of the words that the Awiakay identify as 'spirit language' do indeed strongly resemble Karawari words. In spite of this, we were still unable to produce a meaningful translation of these sections. A possible reason is that as the Awiakay are not fluent in Karawari, the words and phrases borrowed into *Kaunjambi* from this foreign language have become distorted over time. It is therefore impossible, in modern times, to guess what they are supposed to mean in the context of *Kaunjambi.* Parallels to the use of material from neighbouring communities can be found in other New Guinea languages (Voorhoeve 1977: 27; Wassmann 1982; Niles 1991:xii; Rumsey & Niles 2011: 11; Senft 2011), as well as in Australian song languages (cf. Strehlow 1971; Merlan 1987; Curran 2010). (8) and (9) are excerpts from two songs of *Kaunjambi*; the abbreviations used after each line of the song are SL 'spirit language', AA 'Archaic Awiakay' and MA 'Modern Awiakay'. The song numbers refer to the position of the selected text in the sequence of songs. The first line of both songs contain words that even the Awiakay do not understand.

(8) Awiakay, *Kaunjambi* song-cycle, song 2 (excerpt): *Aykmari pambalo*

O-yo, aykmari pambalo, waromaman pambalo (SL)
'???'

yao wambopalik (MA)
'I'm coming up to the house' [repeated many times]

Yao wambopalik, e-e (MA)
'I'm coming up to the house'

etc.

(9) Awiakay, *Kaunjambi* song-cycle, song 42 (excerpt): *Sipin gari*

O, asapi munda konan, ani kunjapoplam (AA+SL)
'O, fiery orb ???'

temun yakandi (MA)
'the sun's about to rise' [repeated many times]

O, asapi koŋganjinbon (AA+MA)
'O, the orb is giving light now'

etc.

The Awiakay are not at all puzzled by the fact that some parts of *Kaunjambi* 'do not have a meaning'. *Aunda endiŋga,* 'just singing', they say. The meanings of magic formulae are just as opaque to ordinary language speakers, and the same can be said of Bible verses, who are only properly understood by prayer leaders and perhaps a handful of other individuals. The spirit language in *Kaunjambi* is simply another such example of an incomprehensible code, which nevertheless serves a key purpose in allowing the song/dance cycle to be performed properly.

As with Eipo and Yale songs, the *Kaunjambi* song texts often refer indirectly to places, events, activities, entities and phenomena that are salient to Awiakay culture. Song 33 (see Example (10)) features an example of using Awiakay terms for colours: *kandukya* 'white', *imbia* 'black', *tolia* 'yellow' and *pawia* 'red' (Hoenigman to appear b). The colour terms are used with the word *mamay* 'stinging nettles'. While the white nettles (*Laportea* sp.) are ones that are distributed widely in the lower parts of the Awiakay land, and the red ones (unidentified sp.) are said to be extremely painful and only found in the mountains (in Awiakay land this is considered to be the area between 70m and 1,200m a.s.l.), the black and yellow ones do not exist at all. Some lead singers explained that these terms were only there for the sake of the structure (TP: *bilong bilasim singsing* 'to decorate the song'), but it is clear that the whole song alludes to a boys' initiation ceremony. The colours mentioned are the colours of the clay used in face painting, whereas young boys who were entering the men's house to be initiated were first beaten with stinging nettles.

(10) Awiakay, *Kaunjambi* song-cycle, song 33 (excerpt)

kandukya mamay
'white stinging nettles' [found in the village]

imbia mamay
'black stinging nettles' [do not exist]

tolia mamay
'yellow stinging nettles' [do not exist]

pawia mamay
'red stinging nettles' [found in the mountain, very painful]

Water bodies on Awiakay land are frequently associated with resident spirits, and named after the vegetation typically growing in and around them. These include the water of thorny sago swamps, the water around fallen sago trunks that contain sago grubs, the water underneath the floating broad-bladed grass, the water around *tendem*, a vine with edible leaves, the water around the roots of pandanus trees, as well as in the water in which *kunai* grass grows (Hoenigman 2015a, to appear b). The variables in this sextuplet are therefore related on the basis of

Awiakay understanding that the spirits can be invoked by calling the names of these places.

(11) Awiakay, *Kaunjambi* song-cycle, song 16 (excerpt)
tao kanja yomba, takia ke
'spirit of thorny sago swamp water, oh, spirit!'
wao kunda yomba, takia ke
'spirit of the water of the sago-grubs trunk roots, oh, spirit'
kongonoŋ kunia yomba, takia ke
'spirit of the broad-bladdered-grass roots water, oh, spirit!'
tendem kunia yomba, tasia ke
'spirit of tendem roots water, oh, spirit!'
kiakay kunia yomba, takia ke
'spirit of pandanus roots water, oh, spirit!'
awe kunia yomba, takia ke
'spirit of kunai grass roots water, oh, spirit!'

7. Offence, criticism and avoidance in Eipo/Yale and Awiakay

Just as there are behaviours that are widely regarded as immoral in Papuan society (hiding food, stealing betel nuts, adultery, going against traditional marriage patterns, and so on) there are also strong social norms against offences caused by the inappropriate use of language. Chief among these are malicious gossip and open public criticism, especially of prominent community members (Hoenigman 2012, 2015b). Critique and open words offend, they require resistance and action, they call forth rage or shame. For this reason, social criticism is made less obvious through ritualized or stylised speech, in song, myths and dancing-song — in order to protect the author or the performer against retaliatory action. We will see which artistic means is good and which style is used against retaliation.

While indirection and disguise of criticism are one means of avoiding retaliation used against authors and performers of songs and dancing-songs, it is a general strategy of narrators of myths and sacred rites in Eipo and Yale. In particular, names of the ancestors, of marsupials and special parts of the body should be strictly avoided. The series of sacred names of the ancestors as well as the names of totems and parts of them cannot be openly mentioned. The narration of myths is kept to dyads like father and son or to pairs of people in special relationships. Songs and dancing-songs can be heard everywhere — on one of my (Heeschen) field trips, Doknoban told myths the whole way from Angguruk to Kosarek, i.e.,

over a period of 10 hours. Songsonga, on the other hand, was always carefully watching the people nearby who were able to listen whenever we stopped.

The song in (12) derides two men who went hunting, and caught only one cuscus, though they had set out for distant mountain areas. They did not bring back any of their catch to share, as they had completely eaten up the cuscus. Moreover, the way they have been moving has been clumsy and awkward; they have been swaying and groping for the ground and fallen branches.

(12) Eipo, *limne*-dancing tune

Neica, dine tektonbo kunmo ara,
'Friend, there has been only one cuscus,'

Iriblin dub bekir, Langkin dub bekir,
'(that is why) you must have gone to the Iriblin top, you must have gone to the Langkin top,'

dine tektonbo kunmo ara, foba dibkiribrobmunge,
'as there was only one cuscus, they have completely eaten up the prize,'

Talinto emai, Cakmanto emai,
'the father of Talinto, the father of Cakmanto,'

tokwe asi talebrin, boke asi talebrin.
'they (have been going) touching the soil, touching leaves.'

The author of the song and the person who actually sings it in the group of dancing men and the onlookers are not identical, and so pride, triumph, and boasting can be openly expressed in the text. However, the power of the dancing songs becomes apparent in a flash when boasting, criticism, or ironic allusion occurs in an inappropriate place and in front of the wrong audience: During an exchange feast, Somsom was so careless as to personally sing the following song that he himself had authored:

(13) Eipo, *sang*-dancing tune, sung by Somsom

A-nerme sukun taren-o,
'This girl, longing, young woman, oh,'

ne kwaning boukwe malinye-uk-cay,
'my longing for sweet potato (i.e. the woman) is at a bad place (she is already married),'

okwas ton yungdiryanurume-e,
'you have cooked and brought my (intense) feeling, eh (you, the husband, have cared for my sweet potato),'

dariryanurum-e.
'you have cooked and brought it, eh.'

Somsom had had an affair with Embemner, the wife of Tirkei. All present knew at once the meaning of the song. The husband was present, and upon hearing the song, took up his bow and arrows. Hosts and guests fought, but luckily there were no casualties, as some women were able to settle the dispute. This event took place in 1977, but the song was still a matter of laughter in 2014.

Ridicule and mockery are pervasive, and one is often warned that the song or even one word or wording of a song shall not be used in the presence of the person criticized or even only mentioned in the song. Thus, mockery is realized in rather old texts, and even dancing-songs will only mention criticized events and protagonists in passing. Perhaps the mockery in old Betengde's song in (14) refers to a by-gone event:

(14) Eipo, *Deren dit* (Mocking song), sung by Betengde

Lim buksalupe, bombom notuma,
Lim PLANT pull moss
'The *buksalupe* moss from Lim mountain, the moss one pulls out,'

Asa daka kalulamkir ciamkine,
forehead split it.covers.you I.have.seen.you.
'I have seen that, having been pulled out, it covers your forehead,'

Lim buksalupe, bombom notuma,
Lim PLANT pull moss
'the *buksalupe* moss of Lim mountain, the moss one pulls out,'

seka kalulamkir eiamkine
cut.off it.covers.you I.have.seen.you.
'I have seen that, having been cut off, it covers you,'

boana dongkune, tanar dongkune,
carry I.will.fetch wasp I.will.fetch
'I will bring you the goods, I will bring you the wasp.'

Sungkule burumburum yin beiande,
bird.arrow shiver bow bring
'Bring the bird-hunting arrow, the shivering bow,'

mar burumburum yin beiande
arrow shiver bow bring
'bring the arrow for the shivering bow.'

The text refers to a married man who felt a deep craving for a vagina during hunting. In many New Guinea cultures having sex before hunting is taboo, as nearby spirits can smell bodily fluids. If a hunter were to have sex shortly before setting out to hunt, the spirits would cover the game, so that he would not see it. On a more practical note, he might also become accident prone, as he would be dis-

tracted by his desire. In the song, the hunter can only think about the moss he wants to grasp, and about the birds that he wants to catch. In other words, he can only think about the woman and her vagina. 'The goods' in line 5 (lit. 'carrying') could refer to a hunting dog who fetches prey, as well as a woman who delivers a netbag with sweet potatoes for the hunter. The 'wasp', on the other hand, refers to 'performing unseemly activities'. The 'shivering bow' is a small one, but the word *burumburum* (glossed as 'shiver' above) can also mean 'be in heat'. The singer is therefore saying to the hunter: "You know nothing of hunting, so go home and hide in a vagina!"

While the Awiakay do not criticise through song per se, they have equally sophisticated spoken means of airing their grievances (Hoenigman to appear b). As mentioned earlier, open public criticism or shaming is generally frowned upon, but it does occur when the injured party feels particularly aggrieved, and/or feels that s/he has a sympathetic audience. Even the public criticism of a 'big man' or community leader can take place under the right conditions, such as when the injured party is possessed by a powerful spirit; in one recorded case, this was a Christian spirit — the Virgin Mary, or, as the Awiakay call her, Mama Maria (Hoenigman 2012). Here, the woman possessed by Mama Maria used a number of linguistic and extra-linguistic devices, such as rhetorical questions, addressing the audience, laughing at and denying what was said by her target (a prayer leader) at key points in the quarrel, and perhaps most importantly, by always talking in Awiakay, even though the prayer leader chose to counter in English or TP. English and TP are generally regarded as the languages of power and authority, but Awiakay is not completely powerless: its implicit potency lies in the fact that it gives its speakers a feeling of safety and intimacy. People tend to resort to Awiakay whenever they feel threatened by the unfamiliar, whether in the form of people or places. Awiakay is an indispensable tool for hiding one's intentions and misleading the enemy, as well as for insulting and challenging others (cf. also Kulick 1993).

When it comes to the content of what is actually said, different expressions have different levels of ability to offend, shame and provoke the target person. These range from mildly abusive language (referring to bodily imperfections or low intelligence), medium-strong expressions (referring to male genitals, plus borrowed swear words from TP), very strong abusive expressions (references to male defecation, female genitalia and menstruation) and strict taboo expressions (accusing someone of being a *sanguma*, 'assault sorcerer' in TP). When tempers flare, of course, many of these expressions do get used. In the following public domestic quarrel, Namang, a woman who had been nagging her husband Wanayk for being lazy, and was then hit by him, flew into a rage, and went on to shame him by accusing him of sodomy, vividly describing his socially unacceptable sex-

ual conduct towards her, and pointing up the lack of reciprocity in the situation by repeating how none of her brothers were doing the same to his sister (Hoenigman 2015b). The last point was made to tell the gathered audience that Namang's marriage was 'one-sided', and that her brothers had not received Wanayk's sister as an 'exchange woman'. Namang and Wanayk's marriage had therefore not been carried out properly.

(15) Awiakay, the shaming of Wanayk, by Namang

Anamgoy nambokonjia tuŋgumgon kay aka kumbokopla.
'My brothers don't stuff his sister's hole.'

Tuŋgumgon kopoñikmun anamgoy nambokonja aka ekua koŋ pokondimbon.
'My brothers didn't marry his sister so that he would beat me like this.'

An tay inja wansait inja kombokondimbon popon anamgoy kay kombokopla toŋgoy kopoñikmun.
'He thinks it's all right to stuff my hole 'one-sidedly' [without exchange], while my brothers aren't screwing his sister.'

Anamgoy nambokonjia toŋgoy kopoñikmun kay aka warikepla anamgoy nambokonja mamgoy injua.
'My brothers aren't probing his sister's cunt with hands.'

Aka aunda!
'Not at all!'

Nan kakape toŋgoy kopoñikmun nangoy nambokonja mamgoy injua kay warikiplam?
'Tell me, are my brothers probing your sister's cunt?'

Agalon! Agalon!
'No! No!'

A final verbal device that can be considered in this section is the Awiakay practice of *pukupuku kaŋapla*, which is best approximated by the term coined by Steven Feld: 'sung-texted melodic weeping' (Feld 1982: 93). Although primarily expressing an individual person's grief at the death of a loved one, Awiakay laments have an additional function: they are a legitimate means for pointing at existing imbalances in the community and thus play an important part in the process of righting such perceived wrongs (Hoenigman 2015a). Melodic weeping or lamenting invariably deeply affects people who hear it and triggers sympathy in them. The person who weeps in this way while publicly making subtle criticisms is not being judged. If this occurs in the house of the deceased, people show the lamenter their respect by seeming not to be disturbed by it, but quietly carrying on with whatever they are doing.

Following the death of his brother Akaim, the prayer leader Moyambe (mentioned above in the Mama Maria incident) hinted during his lament in (16) that no one other than Akaim had ever helped him (e.g. by providing food, or through work on his house) in spite of the good work and sacrifices he had made for the community.

(16) Awiakay, lament for Akaim, by Moyambe

Wok misin ambla aka mokoym-mokoym-pisip popalik, ayo, Akaim o!
'I don't think I can do my mission work by myself, *ayo*, Akaim oh!'

...

Akaim o, nan ambla lukautimbapundiman wok misin mokopokoy.
'Akaim oh, you alone were taking care of me when I was doing my mission work.'

Laments can also be performed for dogs, who are regarded as valuable hunting companions (Hoenigman 2015a). In the days when the Awiakay still practiced male initiation, one of the most important things that a boy learned during the course of it was how to use hunting magic to make his dog kill many pigs. Apart from people and spirits, only dogs and spirit objects have names. Not only among the Awiakay, but in the wider Sepik area, a puppy is an important gift which eventually needs to be reciprocated. A dog becomes an extension of its owner[2] and is, in Awiakay belief, the only animal that can, like humans, be attacked and killed by an *emay* 'assault sorcerer' (TP *sanguma*).

(17) Awiakay, lament for a dog, by Kununda

Tam — o, e-e.
'Dog — oh, e-e.'

Maninjanda paŋaŋgombek olukunja opiaŋgombep kopa maŋgombalik paŋgombuk aka timbupua.
'I saw people [from Konmei who gave me the dog] and I asked them [my sons] for some money to reciprocate the gift, but they didn't give me any.'

Aunda aŋgainim yambiaŋ, yambiaŋ, yambiaŋ, ya meŋgenja aunda pokopep nombopua, aunda maŋgombok.
'I've been walking around in deep worry, while they've been eating the meat my dog had killed. I gave it to them for nothing.'

Olukunja isoŋ yambakay iskamianjanda aka timbupua.
'They wouldn't even give me so much as a bit of eaglewood to sell.'

2. In Ambonwari, for instance, every activity in Ambonwari is extended to the things which are used in that activity (Telban 1998: 162–163, 174); just like Awiakay hunting dogs, the identities of things are determined by the people who use them.

Kenda pokaŋguŋga pekepua kumbiŋ.
'They'd think nothing of hitting the dog when they came down from the mountain.'

Kenda meŋgea nombopua, kenda pokaŋgumbua.
'They'd eat meat again and they'd hit my dog again.'

In her lament in (17), Kununda expresses her grievance that her sons would not help her reciprocate for the now deceased dog that had been given to her by people from Konmei village. By failing to do so they prevented her from being a moral person and fulfilling her obligation. While Kununda's sons had all benefited from the dog, they did not pay attention to her worries that she was indebted to people from another village. In this way, they had wronged her. Moreover, by hitting the dog (probably chasing him away while they were eating his kill) they wronged the dog as well, which was again directed at Kununda, as dogs are their owners' extensions.

8. Conclusion: Engaging with song texts

Engagement with song texts necessarily goes through multiple phases (first encounter, documentation, transcription, translation, commentary), and in each phase, the interplay of philology, demystification and interpretation is evident. This is made especially clear in the above translations of the Eipo and Yale songs, which overflow with indirection and allusion, and require additional commentary to allow outsiders to understand the true intent of the texts. Translation, therefore, along with the inherent reflexivity of the languages involved, should mirror the incompleteness suggested by the unalterable differences of speaker versus hearer, versus reader, observed language versus language of description, and language versus other symbolic codes. Linguists and anthropologists alike as translators could, at least momentarily, refrain from theory in order to remain close to the data or give an account of how the data are discovered and constructed. Perhaps s/he will look at translation as a means of (philological) delineation, of defining the task specificity of speech in different cultures and of bringing gaps within and between speech communities. As for the anthropologist, translation develops into a metacode pointing to gaps, balancing differences, and requiring devices that mediate between observations in the field and experiences transposed into theory.

An unexpected practical challenge was encountered while attempting to translate *Kaunjambi* songs into English for publication. There are several cases in the *Kaunjambi* text, in which variables in parallel sets are a combination of (near-)synonyms and words denoting the same referent in different codes (including neigh-

bouring languages, Archaic Awiakay and spirit language). One prominent example is Song 24, where all the variables in a parallel set mean '(spirit) crocodile', appearing in the order: ordinary Awiakay — synonym A — Miyak borrowing — archaic Awiakay — synonym B (Hoenigman 2015a, to appear b), see (18). To stay as close to the original as possible and to convey the feeling of Awiakay speakers, and at the same time preserve intelligibility for English speakers, I chose to translate the archaic Awiakay *awai* 'crocodile' with the Middle English *cocodrille*, and the Miyak borrowing *asin* with the Gaelic *crogall*. In the translation, these terms appear in italics.

(18) Awiakay, *Kaunjambi* song-cycle, song 24 (excerpt)

kaman epop, papaka-imbuk-e
'I thought you were a crocodile and I went and threw you [into the water]'

tasiam epop, papaka-imbuk-e
'I thought you were a *tasia*-spirit and I went and threw you [into the water]'

asin epop, papaka-imbuk-e
'I thought you were a *crogall* and I went and threw you [into the water]'

awai epop, papaka-imbuk-e
'I thought you were a *cocodrille* and I went and threw you [into the water]'

Papuap epop, papaka-imbuk-e
'I thought you were the *spirit Papuap* and I went and threw you [into the water]'

It should be apparent to the reader that the text of the songs presented in this chapter would not serve as good source material for the grammars of any of the languages that the songs are supposed to be in. Many words and expressions used in the songs have their origins in different dialects or languages, as well as diverse stages of history of the languages. Individual native speaker consultants may sometimes know the history and author of the song, and therefore the origin and import of any unusual words contained in it, but often people merely say *nun walwal, kelape sik tenine* 'we don't know, that is the women's thinking'. This statement is reminiscent of the Awiakay's total acceptance of archaic or unintelligible sections in the *Kaunjambi* songs, and the Meakambut's insistence that songs should only be composed in other languages, and not their own. But talking about the contents helped to uncover new words and locally salient categories and collocations, yielded deep insights into a community's *Weltanschauung*, and above all, provided reasons to laugh and continue asking new questions.

References

Curran, Georgia. 2010. "Linguistic Imagery in Warlpiri Songs: Some Examples of Metaphors, Metonymy and Image-schemata in Minamina Yawulyu." *Australian Journal of Linguistics* 30 (1): 105–115.

Feld, Steven. 1982. *Sound and Sentiment: Birds, Weeping, Poetics, and Song in Kaluli Expression* (2nd ed. 1990). Philadelphia: University of Pennsylvania Press.

Foley, William A. 1986. *The Papuan Languages of New Guinea*. Cambridge: Cambridge University Press.

Franchetto, Bruna. 2006. "Ethnography in Language Documentation." In *Essentials of language documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 183–211. Berlin: Mouton de Gruyter.

Garner, Mark W. J. 2005. "Language Ecology as Linguistic Theory." *Kajian Linguistik dan Sastra* 17 (33): 91–101.

Godschalk, Jan A. 1993. *Sela Valley: An Ethnography of a Mek Society in the Eastern Highlands, Irian Jaya, Indonesia*. Amsterdam: Academisch Proefschrift Vrije Universiteit.

Heeschen, Volker. 1990. *Ninye Bún. Mythen, Erzählungen, Lieder und Märchen der Eipo.* Berlin: Reimer.

Heeschen, Volker. 1998. *An Ethnographic Grammar of the Eipo Language Spoken in the Central Mountains of Irian Jaya (West New Guinea)*, Indonesia. Berlin: Reimer.

Heeschen, Volker. 2005. "Emphatic Speech, Expressive Genres, and the Dancing-Songs of the Eipo and Yalenang (Eastern Mountains of West Papua)." In *Expressive Genres and Historical Change. Indonesia, Papua New Guinea and Taiwan*, ed. by Pamela J. Stewart, and Andrew Strathern, 235–258. Hants/Burlington: Ashgate Publishing.

Heeschen, Vollker, in Zusammenarbeit mit Yakob Bolmerin, Laik Malyo, Eneneas Malyo, Enus Nabyal und Filipus Nabyal. 2015. *Die Eipo in Papua. Weltbilder, Ethnographie und Erzählungen*. München: Institut für Ethnologie.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2006. "Language Documentation: What is it and What is it Good for?" In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin: Mouton de Gruyter.

Hoenigman, Darja. 2007. Language and Myth in Kanjimei, East Sepik Province, Papua New Guinea. M.A. Thesis, Ljubljana: Graduate School of the Humanities.

Hoenigman, Darja. 2012. "A Battle of Languages: Spirit Possession and Changing Linguistic Ideologies in a Sepik Society, Papua New Guinea." In *Interlingual Articulations in Asia and the Pacific: Figuring Sociocultural Otherness Through Otherness of Linguistic Codes*, ed. by Alan Rumsey, and Rupert Stasch, 290–317.

Hoenigman, D. 2015a. 'The Talk Goes Many Ways': Registers of Language and Modes of Performance in Kanjimei, East Sepik Province, Papua New Guinea. Doctoral Dissertation: The Australian National University.

Hoenigman, D. 2015b. "'Are My Brothers Fucking Your Sister?' Shaming and Being (A)shamed in a Sepik Society." In *Language, Morality and the Emotions*, ed. by Bree Blakeman and Ian Keen, 381–397.

Hoenigman, Darja. To Appear a. "Awiakay (Arafundi)." In *Oxford Guide to the Papuan Languages*, ed. by Nicholas Evans, and Sebastian Fedden. Oxford: Oxford University Press.

Hoenigman, Darja. To Appear b. *Talk Goes Many Ways*. Amsterdam: John Benjamins.

Jakobson, Roman. 1966. "Grammatical Parallelism and its Russian Facet." *Language* 42 (2):398–429.

Kulick, Don. 1993. "Speaking as a Woman: Structure and Gender in Domestic Arguments in a New Guinea Village." *Cultural Anthropology* 8 (4): 510–541.

Laycock, Donald C. 1973. *Sepik Languages — Checklist and Preliminary Classification*. Canberra: Australian National University.

Lowth, Robert. 1779. Isaiah. A New Translation with a Preliminary Dissertation and Notes. critical, philological, and explanatory. London: J. Nichols.

Merlan, Francesca. 1987. "Catfish and Alligator: Totemic Songs of the Western Roper River, Northern Territory." In *Songs of Aboriginal Australia*, ed. by Margaret Clunies Ross, Tamsin Donaldson, and Stephen A. Wild, 143–167. Sydney: Oceania Publications.

Mühlhäusler, Peter. 1996. *Linguistic Ecology: Linguistic Change and Language Imperialism in the Pacific Region*. London: Routledge.

Niles, Don. 1991. "Editor's Introduction." In *The Song to the Flying Fox: The Public and Esoteric Knowledge of the Important Men of Kandingei about Totemic Songs, Names and Knotted Cords (Middle Sepik, Papua New Guinea)*, by Jürg Wassmann, x–xv. Boroko: Cultural Studies Division, National Research Institute.

Rumsey, Alan. 2007. "Musical, Poetic, and Linguistic Form in Tom Yaya Sung Narratives from Papua New Guinea." *Anthropological Linguistics* 49 (3/4): 235–82. https://www.jstor.org/stable/27667610

Rumsey, Alan, and Don Niles (eds). 2011. *Sung Tales from the Papua New Guinea Highlands: Studies in Form, Meaning, and Sociocultural Context*. Canberra: ANU Press.

Rumsey, Alan, and Don Niles. To Appear. "Language, Song, and Sung Tales in the Papuan Region." In *The Oxford Guide to the Papuan Languages*, ed. by Nicholas Evans, and Sebastian Fedden. Oxford: Oxford University Press.

Schiefenhövel, Wulf. 1976. "Die Eipo-Leute des Berglands von Indonesisch-Neuguinea." *Homo: Journal of Comparative Human Biology* 26, 263–275.

Senft, Gunther. 2011. *The Tuma Underworld of Love: Erotic and Other Narrative Songs of the Trobriand Islanders and Their Spirits of the Dead*. Amsterdam: John Benjamins.

Strehlow, Theodor G.H. 1971. *Songs of Central Australia*. Sydney: Angus and Robertson.

Telban, Borut. 1998. *Dancing Through Time: A Sepik Cosmology*. Oxford: Oxford University Press.

Toyoda, Yukio. 1998. "Introduction: Fringe Area of Highlands in Papua New Guinea." In *Fringe Area of Highlands in Papua New Guinea* (Senri Ethnological Studies No. 47), ed. by Shuji Yoshida, and Yukio Toyoda, 1–20. Osaka: National Museum of Ethnology.

Turpin, Myfany, and Tonya Stebbins. 2010. "The Language of Song: Some Recent Approaches in Description and Analysis." *Australian Journal of Linguistics* 30 (1): 1–17.

Voorhoeve, Clemens L. 1977. "Ta-Poman: Metaphorical Use of Words and Poetic Vocabulary in Asmat Songs." In *Language, Culture, Society and the Modern World* (New Guinea Area Languages and Language Study, Vol. 3), ed. by Stephen A. Wurm, 19–38. Canberra: Pacific Linguistics.

Wassmann, Jürg. 1982. *Der Gesang an den Fliegenden Hund*. Basel: Ethnologisches Seminar der Universität und Museum der Kulturen.

Wurm, Stephen A., and Shirô Hattori (eds). 1981. *Language Atlas of the Pacific Area. Part I: New Guinea Area, Oceania, Australia*. Canberra: The Australian Academy of Humanities in collaboration with the Japan Academy.

CHAPTER 14

Transcription practices and (the documentation of) metalinguistic knowledge

Katherine Walker[1], Birgit Hellwig[1] & Dagmar Jung[2]
[1] University of Cologne | [2] University of Zurich

Transcripts are an integral part of documentary corpora, and transcription, conducted in close cooperation with a language's speakers/signers, is central to fieldwork activities. Yet, transcription practices — involving decisions that reduce the complexity of an audio/video signal into a written form — remain surprisingly under-researched within the language documentation context. This contribution investigates variability and agreement among transcribers across three languages (Iha, Qaqet, Dëne Sųłıné), focusing on their phonemic and phonetic spellings of nasalization, and asks to what extent the observed transcription strategies provide a window into metalinguistic knowledge. The results paint a complex picture, partly corroborating the view that speakers are aware of phonemes, but also demonstrating that speakers are aware of considerable phonetic detail and automatic (morpho-)phonological processes.

Keywords: transcription, nasalization, orthography, field methods, Papuan languages, Dene languages

1. Introduction

Transcription is a key component of language documentation, both as a product ("transcript") and a practice ("transcribing"). As a product, a transcript is a written form of an audio/video recording, which serves as primary data for much linguistic analysis. Transcribing as a practice is the process of interpreting the audio/video signal and reducing it to a written form.[1] There is increasing awareness that this process involves analytical decisions — with the result that different tran-

1. We focus on written transcripts, but the documentary toolkit also provides for oral transcripts (e.g. Reiman 2010).

https://doi.org/10.1075/slcs.240.14wal
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

scribers are likely to produce different transcripts. Despite this, there is surprisingly little research on transcription practices within the language documentation context: What decisions do transcribers make? What do these decisions reveal about their metalinguistic knowledge? What is the extent and limit of variation, and how can we account for it? How do variable decisions impact subsequent linguistic analyses? While the documentarist turn has succeeded in shifting our attention to spoken/signed language in natural(istic) settings, we will only be able to fully harness the potential of this data type by shifting our attention to "the core of the transcription challenge: reaching a better understanding of the transcription process itself and its relevance for linguistic theory" (Himmelmann 2018: 35). Drawing on the authors' experiences of documenting three endangered languages (Iha, Qaqet, Dëne Sųłné), this chapter takes a first step towards this goal. By analysing the resulting transcripts, we investigate how speakers approach the task of transcription, asking what these approaches reveal about language structure; that is, to what extent the observed transcription strategies provide a window into metalinguistic knowledge.

Theoretical treatments of transcription as a product date back at least to Ochs (1979), who discusses how to represent which data (or not) in a transcript. This tradition is largely taken up by the field of discourse studies. Additionally, several contributions discuss various facets of transcription as practice (Bucholtz 2000, 2007; Dobrin 2021; Haviland 1996; Jung & Himmelmann 2011). Also relevant here is literature on the relationship between written and spoken/signed language, and editing practices, whereby transcripts based on a recording are altered in order to produce a written product, often for purposes such as community materials (De Vries 2015; Maas 2009; Mosel 2006, 2015).

Transcription as practice can be approached from a practical point of view: What steps must be taken to get from a recording of a spoken/signed communicative event to a written transcript? This approach is adopted in literature on field methods, though usually with a narrow scope, focusing on phonetic aspects of transcription (Jung & Himmelmann 2011: 201; exceptions include Chelliah & de Reuse 2011: 434–435; Crowley 2007: 137–141). The theoretical approach aims instead to answer the question: "What do speakers and researchers actually do when they transcribe (or assist in transcribing)?" (Himmelmann 2018: 35). This question refers not simply to the mechanics of translating from one modality to another, but to the particular metalinguistic skills that come into play during the process of transcription. These skills belong to "a reflective mode (as opposed to a production mode) in dealing with language" (Himmelmann 2018: 36). In practice, practical and theoretical considerations are of course intertwined: how and where a transcript is created will constrain how, which, and whose metalinguistic skills are employed, and how such metalinguistic knowledge can be captured. Impor-

tant considerations include the community's previous experience of transcription (e.g. knowledge of writing systems and standards established by missionaries or educational institutions), and the development of an individual transcriber's transcription system over time, according to discussions with linguists and other local collaborators.

Central to our contribution is the question to what extent variation in transcripts can provide insights into the metalinguistic knowledge of transcribers. We approach this question by investigating the spelling of nasalization in transcripts produced by Iha, Qaqet, and Dëne Sųłiné transcribers, treating spelling variation as a source of data in its own right, rather than dismissing it as "inconsistency" (Dobrin 2021). Focusing on the topic of spelling means that our investigation is of relevance to proposals arguing that speakers are aware of phonemes rather than phones (see Section 3). To date, few studies systematically investigate this proposal: evidence is mostly anecdotal or comes from attitudes to (more or less standardized) orthographies that are disseminated through teaching. Systematic evidence from spontaneous transcription of spoken/signed language, by contrast, is lacking. As such, our study aims to contribute to the evidential base: it identifies transcribers' strategies for rendering nasalization, explores the extent and limit of variation in those strategies, discusses them in light of language-external (especially, exposure to writing practice) and language-internal (especially, language-specific phonologies) factors, and offers hypotheses for future research. The results paint a complex picture, in that they partly corroborate the view that speakers are aware of phonemes, but also demonstrate that speakers are aware of considerable phonetic detail and automatic (morpho-)phonological processes.

While our study focuses on accessing metalinguistic knowledge, the findings have implications for analytical and applied issues. To the extent that they indicate sources of variability, they minimize the danger that transcripts potentially skew our subsequent analyses. And to the extent that they reveal insights into speakers' metalinguistic knowledge, evidence from transcription can inform linguistic analyses as well as practical orthography development, thus potentially enriching descriptive and documentary practice.

This chapter is structured as follows: Section 2 outlines our transcription methodologies, and introduces language-external variables that may have impacted transcribers' choices; Section 3 introduces and motivates our focus on nasalization; Sections 4.1 to 4.3 present the three case studies on Iha, Qaqet, and Dëne Sųłiné; and Section 5 brings the findings together.

2. Methodologies

Ideally, we would compare transcripts produced using the same methodology. However, while all the transcripts analysed here share a common methodological core and format (typed transcripts in ELAN), they originate from separate documentation projects, where different priorities and settings led to differing transcription procedures. Although best practice recommends documenting how transcripts originated, keeping original transcripts, and describing any subsequent modifications, this is not (yet) common practice, and methodological discussions are underdeveloped. As such, an ancillary objective of this chapter is to showcase possible approaches and highlight a number of variables with the potential to influence transcription practice.

The major differences between the three documentation projects and the resulting transcripts are as follows. The documentation of Iha (ihaa1241), a Greater West Bomberai language with up to 6,000 speakers in West Papua, Indonesia, has long been supported by Nikolaus Himmelmann, who has been instrumental in creating the published corpus (Narfafan & Tuturop 2009–2016) and grammar sketch (Walker & Himmelmann, to appear). The Iha transcripts in the case study were collected as part of a preliminary investigation into transcription variation: three native-speaking collaborators each produced a transcript of the same recording, a traditional folktale narrated by a single adult speaker (4 min 50s). The Qaqet and Dëne Sųłiné transcripts were not collected with a view to studying transcription strategies, but were produced in the process of constructing child language corpora that feature spontaneous interaction. For Qaqet (qaqe1238), a Baining language spoken by around 15,000 people in East New Britain Province, Papua New Guinea, we selected recordings (totalling around 60 minutes) that were accidentally transcribed twice by the same two transcribers operating as a team (between 9 and 11 months apart). These were drawn from our longitudinal corpus (Hellwig et al. 2023). Finally, for Dëne Sųłiné (chip1261), a Dene language of Canada, we selected transcripts of different recordings that were produced at different points in time (three transcribers, with at least two transcripts each).

To summarize, each case study has a different basis for comparison: transcripts of the same recording by different transcribers in different transcription scenarios (Iha); transcripts of the same recording by the same transcribers at different points in time (Qaqet); and transcripts of different recordings by different transcribers at different points in time (Dëne Sųłiné). Each case study therefore permits different types of investigations into inter- or intra-transcriber variability, allowing us to identify predominant and minor strategies in the transcription of particular sounds. Across case studies, we can compare the strategies used in each

language. Our contribution focuses on the latter, but the variability within each case study informs our interpretation of the results.

As for the transcription procedures themselves, transcribers in all three languages were asked to produce an "orthographic transcription" (Schultze-Berndt 2006: 220–222) with a free translation into a language of wider communication (Iha to Indonesian, Qaqet to Tok Pisin, Dëne Sųłné to English). However, transcribers did not receive specific instructions concerning orthography and/or representing phenomena of spoken language. Qaqet and Dëne Sųłné transcribers were asked to faithfully render non-adult-like expressions uttered by children, and, where applicable, to additionally produce an edited transcript containing an adult/standard equivalent. This was done to varying degrees, but in the case of non-adult-like pronunciations, transcribers interestingly tended to write the expected adult pronunciation, not the actual pronunciation. We have therefore decided to include transcriptions of non-adult-like pronunciations in our study.

Additional variables are non-exhaustively listed in Table 1 (variables of transcription methodology) and Table 2 (transcriber profiles). Note that a "transcriber" is not necessarily a native speaker of the language in question, but happened to be so in our three case studies.

Table 1. Transcription methodologies (using ELAN)

Language & transcriber	Transcription methodology	Segmented by	Typed by
Iha			
FW	collaborative	R	R
PB	independent	R	self
ET	independent	self	self
Qaqet			
APA/BLN	independent (team)	RA	RA
Dëne Sųłné			
TR1–TR3	independent	RA	self

R: researcher; RA: student research assistant; self: the transcriber(s)

In Table 1, the overall "transcription methodology" is classified as "collaborative", in which the researcher has considerable involvement, or "independent", in which transcribers work alone (or in a team for Qaqet) with varying levels of training and feedback (see Jung & Himmelmann 2011: 206–207). Further variables within these broad classifications include who segments the speech signal into chunks in ELAN and who enters the transcript into the ELAN segments. It is the data entered into ELAN for the first time that is compared here. Depend-

ing on each project's goals, further steps were undertaken by the researchers and/or RAs, such as subdividing segments into intonation units (IUs); checking and completing transcripts and translations (including adding spoken-language phenomena such as hesitations, false starts, etc.); and/or adding further annotation layers. All original transcripts are preserved in both their original formats and on dedicated tiers within ELAN files.

Segmentation in ELAN was done predominantly by the researchers and/or RAs. They tended towards longer segments potentially containing multiple IUs. This had both practical and theoretical motivation: working with fewer segments means less delay or truncation of video playback, avoids introducing potentially artificial breaks (as prosodic phrasing is not yet well understood in the languages), and allows the transcriber to gain an understanding of a larger chunk before writing it down (which prioritises understanding over auditory perception). Only ET performed the segmentation step himself, orientated towards IUs. Typing was done by the transcribers (two of the Iha transcribers; all of the Dëne Sųłıné transcribers), or by either the researcher (FW, Iha) or RAs working from handwritten transcripts (Qaqet).

The consequences of these differences are, first, that we cannot investigate transcribers' segmentation strategies, since most did not perform segmentation themselves. Second, though longer chunks facilitate understanding by providing more context, it means having more material to type (and remember) at a time. Concerning who typed the transcript, potential sources of variation include: the orthographic and analytic biases of the researcher, difficulties experienced by the RAs in deciphering the handwritten transcripts and the introduction of possible typos, ease of typing special characters, and different levels of familiarity with different orthographies among transcribers themselves (see below).

Table 2 gives several relevant variables related to transcriber profiles. Concerning familiarity with the speaker, ET is related to the Iha narrator and is in the video himself (providing backchannel responses). FW was familiar with the tale prior to seeing the video, but did not know the speaker. PB was unfamiliar with the story and the speaker, and comes from a different dialect area. The Qaqet and Dëne Sųłıné transcribers are close relatives of the children in the recordings (mostly parents), and thus very familiar with all participants, though not necessarily with the contents of the recordings. On the one hand, greater familiarity with the participants and/or contents of a recording may facilitate understanding; on the other hand, transcribers may be more likely to make assumptions about what *should* be in a particular utterance, rather than transcribing what they actually hear.

All transcribers used a pre-existing practical orthography, but had variable prior exposure to literacy. The Iha transcribers received BA-level, Indonesian-

Table 2. Transcriber profiles

Transcriber	Age (in yrs), gender, education	Familiar with speaker(s)	Prior experience with…	
			Transcription	Orthography
Iha				
FW	50, M, BA	no	None	Some
PB	28, M, BA	no	None	Likely low
ET	37, M, BA	yes	Several years	High
Qaqet				
APA	34–36, M, High School	yes	None → some*	Some
BLN	31–33, F, High School	yes	None → some	Some
Dëne Sųłiné				
TR1	31–34, F, High School	yes	None → some	Some
TR2	30–32, F, High School	yes	None → some	Some
TR3	30–32, F, High School	yes	None → some	Some

* "None" when producing earlier transcripts; "some" when producing later transcripts.

language education, but varied in their experience with Iha orthography. PB likely had very little, while FW sometimes consults the community Indonesian-Iha dictionary (Widodo et al. 2012). ET has been transcribing Iha using ELAN for several years at the *Center for Endangered Languages Documentation* (see Sawaki et al., this volume) — with instruction and feedback not least from Nikolaus Himmelmann — and helped develop the researcher orthography. The Qaqet transcribers completed High School and were familiar with the basic principles of the practical orthography, but have very little opportunity to read or write Qaqet (or any other language) in their daily lives. We provided some initial training, as well as regular feedback, on technical aspects (e.g. operating ELAN) and general transcription principles (i.e. not writing any words that were not actually uttered), but not on the orthography itself. The Dëne Sųłiné transcribers completed High School, are literate in English, and started with basic knowledge of writing Dëne Sųłiné. They consulted the researcher on transcription issues that arose during the preparation of the transcript.

3. Nasalization

To ensure coherence across our three studies, we have chosen a common focus: the transcription of nasalization, including the transcription of (morpho-)phono-

logical processes that involve nasals and/or nasalized segments. As detailed below, two sets of considerations contributed to this choice.

First, nasalization plays an important role in the phoneme and/or phone inventories of all three languages (and language families), albeit in different ways. Iha and Qaqet both have prenasalized voiced plosives. These are phonemes in Iha (contrasting with oral voiced and voiceless plosives), but phonetic realizations of voiced plosives in Qaqet (contrasting with voiceless plosives). Both languages additionally have nasal + plosive sequences, as well as variable voicing assimilation in the environment of nasals. Dëne Sųłiné has nasal vowels that phonemically contrast with oral vowels, and that are often realized as nasal consonants before obstruents. Additionally, oral vowels are often nasalized in the environment of nasal consonants.

From a synchronic perspective, descriptions of the three languages often devote considerable energy to the analysis of nasalization. This is due to various, sometimes complex analytical issues, such as differentiating between singleton nasalized phonemes and nasal + consonant/vowel sequences, and/or distinguishing phonemes from assimilatory processes arising in the environment of nasal segments. The language-specific complexities are discussed in Sections 4.1 to 4.3. On an applied level, these analytical issues have potential practical implications for advising communities on establishing orthographies. Specifically, they may speak to the issue of recommending shallow (based on phonetic and allophonic realizations) vs deep (based on underlying phonemes and morphemes) orthographies (e.g. Cahill & Rice 2014; Seifart 2006).

From a diachronic perspective, too, nasalization has received considerable attention. Concerning Iha and Qaqet, prenasalization is widespread across languages of the Pacific region and features prominently in discussions of genealogical and areal relationships. Within the Austronesian family, prenasalized plosives are assumed to constitute an innovation of Proto Oceanic, while non-Oceanic Austronesian languages tend to have nasal + obstruent sequences (e.g. Blust 2013; Ross 1998). In many non-Austronesian (Papuan) language families, "the voiced stops are obligatorily prenasalized" (Foley 1986: 61); however, there is much variability, as shown throughout the language and language family cameos in Palmer (2017) and Evans and Fedden (to appear). For some families, prenasalized plosives can be posited at the proto level, e.g. there are proposals to reconstruct voiceless vs prenasalized voiced plosives for Proto Trans New Guinea (Pawley & Hammarström 2017: 136–137) and Proto Greater West Bomberai (Usher & Schapper 2022) (relevant for Iha). In other cases, this is more doubtful; e.g. Stebbins et al. (2017: 881) consider prenasalized stops to be borrowings from Oceanic languages into the Papuan languages of Island Melanesia (relevant for Qaqet).

Likewise, nasal and nasalized vowels are characteristic properties of Dene languages (Krauss & Leer 1981; Rice 1993). Historically, the two sonorant phonemes **w* and **n* developed differently within the family, giving rise to nasals (*m, n*), plosives (obstruentization into *b, d*), prenasalized voiced plosives (^{m}b, ^{n}d), and/or voiced plosives with nasal release (b^{m}, d^{n}) in different synchronic languages and different environments. In addition, nasalized vowels developed in the environment of nasal segments. Synchronically, oral and nasal vowels contrast in languages such as Dëne Sųłiné, e.g. in verb stems where an imperfective form contrasts with a nasalized perfective form (e.g. *ts'ı* 'make OBJ IPFV' vs *ts'ı̨* 'make OBJ PFV'). Old nasal segments may still surface in certain morphophonological alternations, e.g. compare [kǫ́] 'fire' and [mbe-kón-é] 'his/her fire (3SG.POSS-fire-POSD)' in closely related Dené Dhá (Alberta South Slavey). Here, a stem alternation (nasal vowel vs vowel + nasal sequence) depends on the presence of the vocalic suffix of the possessive construction.

Second, spelling is a domain where we observe great variability. We can illustrate this using our Iha and Qaqet datasets. In both languages, we worked with multiple transcripts of the same recordings, which allows us to compare inter-transcriber (Iha) and intra-transcriber (Qaqet) variability. Tables 3 (Iha) and 4 (Qaqet) give a broad impression of the variation between the transcripts, showing the number and percentage of transcribed segments that match across transcribers (i.e. that are transcribed identically, ignoring differences in punctuation, capitalization, acute accents, open/close vowels, and/or hyphenation). In Iha, ET's transcript was taken as the basis (containing 138 segments),[2] and we conducted a pairwise comparison across the three transcriber pairs (FW–PB, ET–FW, and PB–ET). In this comparison, 87–93% of segments did not match. In Qaqet, the recordings contained 910 segments, around half of which (47%) received non-identical transcriptions. This is a much lower proportion than in Iha, but still substantial. It possibly indicates greater consistency within transcribers vs across transcribers. As mentioned above, the greater familiarity of the Qaqet transcribers with the participants may have played an additional role, as it may have facilitated understanding.

We assign segments that were transcribed non-identically to three broad categories: differences that arise from (i) the addition, substitution, or repetition of morphemes, words, or phrases; (ii) word division; and/or (iii) spelling. Note that any segment may contain multiple differences from the same or different categories. In both languages, the most frequent source of variation by far was spelling (i.e. spelling differences were observed in 86% of non-identically transcribed segments in Iha, 76% in Qaqet). Other differences were present in 32%–61% of

2. We divided the longer segments in the other two transcripts to match ET's 138 segments.

segments, distributed slightly differently across the two languages: division into words (61% in Iha, 32% in Qaqet) and the addition, substitution, or repetition of material (51% in Iha, 38% in Qaqet). As the central source of variation across transcripts and languages, spelling is the focus of this contribution, which we limit to variability in one specific phenomenon — nasalization — to keep the scope of the study manageable.

Table 3. Variability between transcripts: Inter-transcriber variability (Iha)

Segments transcribed twice: 138	**N***	**Percentage***
– identical	9–18	7–13%
– non-identical	120–129	87–93%
Observed differences in non-identical segments**		
– addition, substitution, repetition of material	66	51%
– word division	79	61%
– spelling	111	86%

* Displaying minimum and maximum from pairwise comparisons across three transcriber pairs.
** Does not add up to 129 (100%), as segments may contain multiple differences.

Table 4. Variability between transcripts: Intra-transcriber variability (Qaqet)

Segments transcribed twice: 910	**N**	**Percentage**
– identical	484	53%
– non-identical	426	47%
Observed differences in non-identical segments*		
– addition, substitution, repetition of material	161	38%
– word division	135	32%
– spelling	325	76%

* Does not add up to 426 (100%), as segments may contain multiple differences.

Our guiding hypothesis is that investigating the transcription of nasalization provides insights into the metalinguistic knowledge of transcribers. We focus on nasalization, but we assume that generalizations would be transferable to other cases that draw on transcribers' knowledge of phonology, phonetic realizations, and allophonic variation. Sapir (1949: 46) once remarked that "[i]n the course of many years of experience in the recording and analysis of unwritten languages, American Indian and African, I have come to the practical realization that what the naïve speaker hears is not phonetic elements but phonemes." Sapir famously argues for the psychological reality of phonemes — an idea that has become influ-

ential in orthography development: avoid representing surface realizations, follow speakers' awareness of this psychological unit, and "represent sounds the way native speakers perceive them to sound" (Snider 2014: 28; see also Pike 1947). While there is considerable discussion on the units and processes that speakers may or may not be aware of (e.g. Cahill & Rice 2014), there is comparatively little discussion on how to access this awareness. This is the starting point for our investigation: Can transcribers' choices give us a window into their metalinguistic knowledge? Do transcribers systematically write phonemes? To what extent does their writing reflect phonetic realizations? Can we detect strategies?

4. Case studies

In this section, we discuss the transcription of nasalization by speakers of Iha (Section 4.1), Qaqet (4.2), and Dëne Sųłiné (4.3). This includes nasal phonemes (in Iha and Dëne Sųłiné) as well as oral phonemes that can be realized nasalized and/or appear in the environment of nasal segments (in all languages). This approach allows us to compare strategies for spelling phonemes, phonetic features, and positional variants. We adopt standard bracketing conventions to distinguish between the different levels: /phoneme/, [phone], and <grapheme>. For example, Qaqet /məŋka/ [məŋga] 'tree' is variably spelt <menka> ~ <menga> ~ <maga>. In Iha and Qaqet, we can quantify variability across multiple transcripts, distinguishing between tokens and types. For tokens, we count the number of occurrences of a specific spelling of a specific phoneme in a specific environment (e.g. how often /k/ is spelt <k> vs <g> when following a nasal, plus how often the nasal element is included). For types, we count the number of unique words that exhibit a specific spelling, subsuming both simple words and complex words (i.e. including affixes and/or clitics). For example, Qaqet /mə**ŋk**a/ is spelt eight times with <ng>, corresponding to eight tokens, but one type. This distinction allows us to detect any potential skewing in the results: if a word is particularly frequent in the corpus, a specific spelling pattern may appear frequent when counting tokens, but infrequent when counting types.

4.1 Iha

Iha has three prenasalized plosives /ᵐb, ⁿd̪, ᵑg/, rendered <mb, nd, ngg> in the researcher orthography, contrasting with three voiceless plosives /p, t̪, q/, written <p, t, k>, and two voiced plosives /d̪, ɢ/, written <d, g>. (To narrow the scope, we exclude the marginal phoneme /b/ and the coarticulated consonants /q͡p, ᵑᵐg͡b/.) Table 5 lists the relevant phonemes, their most common phonetic realizations in different environments, and the different ways they are transcribed in the case-study corpus.

Table 5. Prenasalized and simple plosives and their realizations (Iha)

Phonemes	Phonetic realizations		Transcriptions in the corpus
ᵐb, ⁿd̪, ᵑg	word-initial/C+	ᵐb, ⁿd̪, ᵑg ~ ᵐp, ⁿt̪, ᵑq ~ b, d̪, ɢ	mb, nd, ngg ~ mp, nt, ngk ~ b ~ ng ~ ngmb*
	word-internal/V+	m.b, n.d̪, ŋ.g ~ m.p, n.t̪, ŋ.q	mb, nd, ngg ~ mp, nt, ngk ~ ng
	word-final	–	–
d̪, ɢ	word-initial	d̪, ɢ ~ ⁿd̪, ᵑg ~ ʁ	d, g ~ nd ~ t
	word-internal	d̪, ɢ ~ ʁ	d, g*** ~ t
	word-internal: N_	(m.b)**, n.d̪, ŋ.g	nd, ngg ~ nt, ngk
	C+	t̪, (q)	t ~ d
	V+	t̪, q ~ d̪, ɢ ~ ʁ	d, g
	word-final	–	–
p, t̪, q	Elsewhere	p, t̪, q	p, t, k ~ d, g ~ qp* ~ ø
	word-internal: N_	m.p, n.t̪, ŋ.q	mp, nt, ngk ~ p
	V+	(β) ~ t̪, q ~ d̪, ɢ ~ ʁ	t, k ~ d, g***

* Graphemes (researcher orthography) for coarticulated stops: <qp> = /q͡p/; <ngmb> = /ᵑᵐg͡b/.
** Brackets indicate phonemes not found in this position in the corpus.
*** Includes <gh>, a spelling variant for the allophone [ʁ].

Two points to note are, first, the high number of allophones attested for some phonemes. Further research may show that allophonic variation is more orderly than this table suggests. Second, some phonemes share the same allophones, e.g. both /d̪/ and /t̪/ can be realized [d̪] at morpheme boundaries following vowels. In the absence of other evidence, we have relied on native-speaker intuition to determine the underlying phoneme.

The allophones of prenasalized plosives overlap with those of other phonemes (highlighted in grey in Table 5). First, in word-initial position, both prenasalized /ⁿd̪, ᵑg/ and the voiced plosives /d̪, ɢ/ can be realised as either prenasalized [ⁿd̪, ᵑg] or oral [d̪, ɢ]. The reason for the variation is currently unclear, but the nasal portion may be less likely to appear following a nasal-final word. Second, word-internally, prenasalized plosives are generally re-syllabified into a nasal coda and an oral voiced or voiceless plosive onset. Hence, a nasal + homorganic plosive sequence is ambiguous between a single underlying prenasalized phoneme or a sequence of two phonemes. We consider only homorganic sequences, since heterorganic sequences do not generally assimilate.

We therefore compare (i) the spelling of word-initial voiced and prenasalized plosives, and (ii) word-internal nasal + homorganic plosive sequences. In addition, we compare voiced and voiceless plosives in order to investigate transcribers' general sensitivity to voicing. This is relevant to the analysis of prenasalized phonemes because there is variation in the voicing of the plosive portion in all positions in a word. However, only word-internally is there potential ambiguity between prenasalized plosives and sequences, since syllable onsets are constrained: /ⁿd̪/ is a permitted syllable onset; /nd̪/ and /nt̪/ are not.

Table 6 shows how the target phonemes were transcribed in the Iha corpus (dark grey indicates majority patterns; light grey, minority patterns). Due to the high rate of word-division differences in the transcripts (61% of segments; see Table 3), this table does not split the results according to position in a word.

Table 6. Transcription of simple oral and prenasalized plosive phonemes (Iha)

Underlying form		Total N	Transcribed as:				
			Voiced plosive	Voiceless plosive	Nasal + voiced plosive	Nasal + voiceless plosive	Other
1.	**Prenasalized/nasal + homorganic plosive:**						
	tokens	242	<1%	<1%	68%*	30%	1%
	types	114	1%	1%	67%	30%	2%
2.	**Voiced plosives:**						
	tokens	163	89%**	10%	1%	0	0
	types	70	89%	10%	1%	0	0
3.	**Voiceless plosives:**						
	tokens	716	8%	89%	0	0	3%
	types	352	7%	88%	0	0	4%

* Four instances of <ng> are counted as nasal+voiced, equivalent to <ngg> elsewhere.

** One instance of <gh>, the local orthographic rendering of [ʁ], is counted as voiced plosive.

The corpus contained 542 types (excluding Malay loans) and 1,121 tokens. The results for types and tokens are remarkably similar, showing that overrepresentation of a few frequent types is not a concern. The results for all phonemes show that the majority were transcribed with the corresponding grapheme(s) — only 158 tokens were transcribed differently (14.1%). These are shown in Table 7, where results per phoneme (in **bold**) total 100% horizontally, and results per context (in *italics*) total 100% for each phoneme vertically: e.g. **1%** of all prenasalized/nasal + homorganic plosive phonemes are transcribed as voiced plosives ("vd. pl."), *100%* of which are word-initial. Table 8 gives a selection of examples.

Table 7. Deviation in transcripts from the underlying phoneme (Iha)

Underlying form	Total N	Transcribed as:*				
		vd. pl.	vl. pl.	n. + vd. pl.	n. + vl. pl.	co-art. pl.
1. Prenasalized/nasal + homorganic plosive:						
All	**67**	**1%**	**1%**		**94%**	**3%**
word-initial	27	*100%*			*38%*	*100%*
word-internal: V(+)_	30		*100%*		*46%*	
word-internal: C(+)_	10				*16%*	
2. Voiced plosive (/d/ only):						
All	**18**		**94%**	**6%**		
word-initial	15		*82%*	*100%*		
word-internal: V(+)_	3		*18%*			
3. Voiceless plosive:						
All	**73**	**78%**	**21%**			**1%**
word-initial	18	*30%*				*100%*
word-final	13	*4%*	*73%*			
word-internal: V(+)_	42	*67%*	*27%*			
	158					

* Matching underlying phonemes and transcribed graphemes indicate a mismatch in place of articulation. This table excludes underlying word-final consonants that were absent in the transcripts. vd. pl. = voiced plosive, vl. pl. = voiceless plosive, n. = nasal, co-art. pl. = co-articulated plosive.

There were only two mismatches in which an underlying prenasalized phoneme was rendered as a simple plosive or vice versa. In (1a-ii) in Table 8, a word-initial prenasalized plosive is rendered by two transcribers as <mb>, and by one as <b>. Though sometimes nasals are absent in word-initial prenasalized phonemes following a word-final nasal, Example (1a-ii) occurs in IU-initial posi-

Table 8. Transcription of simple oral and prenasalized plosive phonemes and allophones: Examples (Iha)

		Phonemic	Non-phonemic
1.	**Prenasalized/nasal + homorganic plosive**		
a.	Word-initial. Transcribed as:		
i.	nasal + voiceless plosive: [^{n}**d̪**ur] /^{n}**d̪**ur/ 'pig'	<**nd**ur>	<**nt**ur>
ii.	voiced plosive: [^{m}**b**umpaq] /^{m}**b**um=paq/ 'k.o.palm=skin'	<**mb**umpak>	<**b**umpak>
iii.	prenasalized coarticulated plosive: [^{m}**b**roh] /^{m}**b**roh/ '?'	<**mb**roh>	<**ngmb**oroh>
b.	Word-internal V(+)_. Transcribed as:		
i.	nasal + voiceless plosive: [q͡pɛ**ŋg**i] /q͡pɛ**ŋɢ**i/ 'now'	<qpe**ngg**i>	<qpe**ngk**i>
ii.	nasal + voiceless plosive: [^{m}bu**mp**aq] /^{m}bu**m**=**p**aq/ 'k.o.palm=skin'	<mbu**mp**ak>	<mbu**p**ak>
2.	**Voiced plosive**		
a.	Word-initial. Transcribed as:		
i.	nasal + voiced plosive: [**d̪**on**d̪**ihin] /**d̪**on**d̪**ihin/ 'therefore'	<**d**ondihin>	<**nd**ondihin>
ii.	voiceless plosive: [qɔrɔp **t̪**ein] /qɔrɔp **d̪**ein/ 'cousin RPRT'	–	<korop **t**ein>
3.	**Voiceless plosive**		
a.	Word-initial. Transcribed as:		
	voiced plosive: [qo**t̪**naŋqa **ʁe**] /qɔ**t̪**naŋga **qe**/ 'hungry so'	<kɔtnangka **k**e>	<kotnanggag**e**>, <kotnangga **g**eh>
b.	Word-internal V(+)_. Transcribed as:		
	voiced/voiceless plosive (different place of articulation): [^{m}bɾu**d̪**waɾeq] /^{m}bru-**t̪**=waɾeq/ 'run.PL-SF=all'	<mbru**t**warek>	<mbru**d**warek>, <mbru**k** warek>

tion following an approx. 4-second pause. In (2a-i), a simple voiced plosive is rendered twice as <d> and once as <nd>. A third mismatch, in (1b-ii), shows a nasal + voiceless plosive sequence across a morpheme boundary, where the nasal portion is present in two transcripts and missing in the third. The final minor variation for underlying prenasalized/nasal + homorganic plosives is shown in (1a-iii): one transcriber uses the co-articulated prenasalized plosive grapheme <ngmb> in word-initial position, where the other two use <mb>.

Most variation concerns the voicing of a plosive following a homorganic nasal: 95% of mismatches were due to spellings with voiceless plosives rather than voiced (see Table 7). In (1a-i) in Table 8, a word-initial prenasalized phoneme is rendered <nd> by two transcribers and <nt> by the third; in (1b-i) the same pattern is shown for a word-internal sequence of nasal + homorganic plosive. This variation shows a systematic pattern of inter-transcriber difference: of 63 instances of voiceless plosives following nasals, 60 are from ET's transcript. Although almost half of these instances (n=30) are from just two words that include the form /^{n}d̪ur/ 'pig', ET's preference for voiceless plosive graphemes extends to all prenasalized phonemes: he uses <nt, mp, ngk> in any position in the word, e.g. word-internal <ngk> in (1b-i).

Comparing the spelling of simple plosives allows us to investigate sensitivity to voicing in general. As shown in Table 6, 8–10% of simple plosives are transcribed with different voicing than the underlying form. For example, in (3b) in Table 8, the stem-former /t̪/ is written as both <t> and <d> (and <k>, see below). Much of the deviation from the underlying form is systematic, with all three transcribers often representing voicing assimilation through spelling. For instance, the reportative particle /d̪ein/ is consistently rendered with initial <t> following a consonant, shown in (2a-ii), and with initial <d> following a vowel, as in <njo dein> /injɔ́ d̪ein/ 'yes RPRT'. This holds whether the forms are transcribed as independent words or joined to the previous word, and whether the underlying form is voiced or voiceless. That is, (3a) is a rare example of variation: almost all other examples of /qe/ 'but; so' are transcribed with initial <g> following a vowel-final word and initial <k> following a consonant.

Over 20% of variation among underlying voiceless plosives is due to transcribing a different place of articulation (Table 7). Though not directly related to the topic of nasalization, we briefly mention it as further evidence for the value in studying transcription variation. Variation in place of articulation shows a fairly robust inter-transcriber pattern: where FW and ET agree on place of articulation, some of PB's coda consonants assimilate to the place of articulation of the following segment. For instance, in (3b) in Table 8, PB's underlying /t̪/ becomes <k> in <mbruk>, "assimilating" to the velar articulation of the labiovelar /w/ that follows. The equivalent segment is <d> in FW's transcript (showing voicing assimilation) and <t> (showing the underlying phoneme) in ET's.

4.2 Qaqet

The Qaqet case study focuses on prenasalized voiced plosives. Table 9 summarizes all plosive phonemes, their phonetic realizations, and the attested transcriptions. The relevant phonemes and phones are highlighted in grey: voiced plosives (which are often realized prenasalized) and voiceless plosives following nasals (which become voiced in this context).

Table 9. Plosive phonemes and their realizations (Qaqet)

Phonemes	Phonetic realizations	Transcriptions in the corpus
b, d, g	(^{m})b, (^{n})d, (^{ŋ})g	(m)b, (n)d, (n)g
p, t, k	p, t, k	p, t, k
	following nasals:* N+b, N+d, N+g	mp, nt, nk ~ (m)b, (n)d, (n)g
	intervocalic: β, ɾ, ɣ**	p, t, k ~ v, r, q ~ pv, tr(l), kq
	word-final: pˀ, tˀ, kˀ	p, t, k ~ any voiceless plosive ~ Ø

* N = Nasal phonemes /m, n, ɲ, ŋ/. The practical orthography renders /ɲ, ŋ/ as <ng>. Nasal + plosive sequences are not necessarily homorganic.
** Realization varies between fricative/rhotic and approximant; realization of the velar phoneme ranges from uvular to palatal.

Unlike Iha, Qaqet does not have prenasalized phonemes. However, voiced plosives exhibit strong pre-voicing and are frequently realized prenasalized. In a phonetic study of 379 voiced plosive tokens (Tabain et al. 2022; Tabain & Hellwig 2023), 59% were identified as prenasalized. The prenasalized tokens had unusually long closure periods (around 100 ms, including around 60–70 ms of prenasalization), while the non-prenasalized tokens were shorter and more variable (between 30–70 ms), but also displayed strong voicing, suggesting "at least some nasal leakage even in the non-prenasalized tokens" (Tabain & Hellwig 2023: 1079). That is, prenasalization and "nasal leakage" in Qaqet may be interpreted as side-effects of maintaining voicing throughout long closure durations. Since prenasalization is not phonemic, we would expect transcribers not to write it. However, there are indications that speakers are aware of this phonetic feature: they point out prenasalization in elicitation sessions, and the corpus even features spontaneous discussions about this topic.

In a second context, voiced plosives occur as allophones of voiceless plosives following nasals. Within words, there are only very few exceptions to this voicing assimilation: loanwords (e.g. [a**ŋk**əl] 'uncle') and aspectual verb stems (e.g. [u**nt**alak ~ u**nd**alak] /un=talak/ '1DU.SBJ=play.CONT'). Across word boundaries, assimilation is frequent, but modulated by speech rate and careful pronunciation.

In all cases of assimilation, we would expect transcribers to write the nasal element (as it represents a separate nasal segment), while there may be variability with regard to writing the voiceless plosive or its voiced counterpart.

To investigate the above expectations, we identified 1,277 tokens (representing 183 types), and coded their transcription (as voiced plosive, nasal + voiced plosive, or nasal + voiceless plosive). Table 10 summarizes majority (dark grey) and minority (light grey) patterns, and Table 11 exemplifies them.[3]

Table 10. Transcription of voiced plosive phonemes and allophones (Qaqet)

Underlying form	Total N	Transcribed as:		
		voiced plosive	nasal + voiced plosive	nasal + voiceless plosive
1. Voiced plosive:*				
tokens	926	91%	8%	0%
types	92	85%	13%	0%
2. Nasal + voiced plosive (within word):**				
tokens	98	6%	94%	0%
types	13	31%	69%	0%
3. Nasal + voiceless plosive (within word):**				
tokens	194	1%	84%	15%
types	38	3%	82%	16%
4. Nasal # voiceless plosive (across word boundaries):**				
tokens	59	0%	47%	53%
types	40	0%	50%	50%

* Does not add up to 100%, as 5 tokens (2 types) were transcribed differently and are not included in this table.

** Coding reflects transcribers' decisions: if they wrote single words, we coded it as "within word", if they wrote two words, we coded it as "across word boundaries".

Table 10 shows the extent of variability. First, the voiced plosive phoneme is predominantly written without prenasalization (91%), but spellings with a nasal are also attested (8%). The second category covers words that consist of a nasal segment followed by a voiced plosive phoneme. While the nasal segment is usually written (94%), it is interesting to note that we also find cases without the nasal: 6% of tokens, but 31% of types; i.e. the token numbers underestimate

3. We did not separately investigate the spelling of nasal + voiced plosives across word boundaries (analogous to category 4 in Tables 10 and 11), as we did not expect any variability.

Table 11. Transcription of voiced plosive phonemes and allophones: Examples (Qaqet)

	Transcript A	**Transcript B**
1. Voiced plosive, e.g.:		
[mu^{n}**g**un] /mu^{n}**g**un/ 'sit' (LongZDL20160112_1 0069)	<mu**g**un>	<mu**ng**un>
2. Nasal + voiced plosive (within word), e.g.:		
[^{n}giaaɽə**mb**əm] /^{n}gia=aɽə**m**-^{n}**b**əm/ 'your thigh (2SG.POSS=thigh-SG.RCD)' (LongZDL20160202_1 1059)	<gia rla**b**em>	<gia re**mb**em>
3. Nasal + voiceless plosive (within word), e.g:		
[aarə**ng**i] /aa=rən-**k**i/ 'his name (3SG.M.POSS=name-SG.F)' (LongZDL20160202_1 1208)	<arle**ng**i>	<arlen**k**i>
4. Nasal # voiceless plosive (across word boundaries), e.g.:		
[aɽki**ŋ g**ua] /a=ɽki**ŋ k**ua/ 'where are the teeth? (NM=tooth where)' (LongZDL20160202_1 1019)	<arlki**ng g**ua>	<arki**ng k**ua>

the extent to which transcribers omit the nasal.[4] The final two categories cover sequences of nasal segments plus voiceless plosives. In both cases, transcribers invariably write the nasal (with one single exception). Furthermore, they tend to write the voicing assimilation within words (84% vs 15%), as well as in about half of the cases across word boundaries (47% vs 53%).

The four categories thus display markedly different patterns. We can identify a strong preference for not representing phonetic prenasalization (first category). This differs from the other three categories, which predominantly or exclusively include the nasal. And while we notice some blurring between the first two categories, the likelihood of writing the nasal differs between the two. That is, with regard to the nasal element, transcribers tend to write phonemically, not phonetically. In the case of the voicing assimilation (third and fourth category), by contrast, there is a tendency to write the assimilated sound, although this is modulated by its position within or across words. These spelling patterns are consistent across the earlier and later set of transcripts; that is, there was no discernible change over time.

4. The skewing is due to an exceptionally large number of a single type: 45 tokens of <ma**mb**u> 'bamboo' (from Tok Pisin). Interestingly, we also observe spelling variation in such loanwords, e.g. <si**nd**aun> vs <si**d**aun> (from Tok Pisin <si**nd**aun> 'sit down').

4.3 Dëne Sųłiné

Our Dëne Sųłiné case study focuses on vowels. The language has five nasal and six oral vowel phonemes (Cook 2004: 2). Phonetically, the nasalization can either be realized on the vowel itself (i.e. a nasal vowel) or as a sequence of oral vowel + nasal consonant (known as "nasal unpacking"; Paradis & Prunet 2000). The potential realization is largely predictable: if a nasal vowel is followed by an obstruent, we observe nasal unpacking (as in (1)); conversely, in stem-final position (e.g. in a perfective stem), we find the nasal vowel (as in (2)).

(1) Nasal unpacking: [ts'ur**ɪnd**í] /ts'ur**ɻd**í/ 'you (SG) are lazy'

(2) Nasal vowel: [-ts'**ɻ**] /-ts'**ɻ**/ 'make OBJ PFV'

In the practical orthography, nasal vowels are spelt with a nasal hook (<ą, ę, ɨ, ǫ, ų>), while plain graphemes <a, e, ë, ı, o, u> are used for oral vowels. The orthography adopts a strict phonemic spelling and uses hooked vowels regardless of the phonetic realization. That is, even in cases of nasal unpacking, the spelling is, e.g. <ɻC> (not <ɪnC>).

Compared to the Iha and Qaqet transcribers, the Dëne Sųłiné transcribers are experienced users of the national (English) orthography, as much of their school-based education and everyday reading and writing practice is in English. Against this background, graphemes and grapheme combinations that are specific to Dëne Sųłiné orthography are sources of insecurity: this concerns the nasal vowels, but also the spelling of other phonemic distinctions (e.g. tones, the lateral fricative, or the grapheme <ë> for /ɛ/). While untrained transcribers are aware of the existence of hooked vowel graphemes, they are often uncertain about the spelling of nasal phonemes and phones in specific cases, and address these uncertainties in discussions with the researchers. This includes both the spelling of nasal vowels (which may be phonetically oral in cases of nasal unpacking) and nasalized vowels (which may be phonetically nasal but phonemically oral due to assimilatory processes).

As their experience develops, transcribers increasingly resort to the standardized, phonology-based spelling and use the hooked letters. Their initial transcripts, however, show variability, and we can detect the following patterns. In the case of nasal unpacking, untrained transcribers prefer to write the actual realization, i.e. the sequence of oral vowel + nasal + obstruent. This is by far the dominant pattern for labial and alveolar places of articulation (as in (3)). Interestingly, this observation does not hold for velars, where we observe considerable variability. This is illustrated with the help of all attested spellings of /dęges/ 'be quiet!' (sometimes plus additional morphology): these are variably transcribed as oral vowel + obstruent (i.e. without nasalization; (4a)), nasal vowel + obstruent (i.e.

with the nasal hook; (4b)), and oral vowel + nasal + obstruent (i.e. with nasal unpacking; (4c)). The patterns in (3) and (4) suggest that transcribers are aware of the place of articulation of the nasal segment. This is unproblematic in the case of labial and alveolar nasals (which are easily rendered as <m> and <n> respectively), but necessitates finding ad hoc solutions for dealing with the lacking velar nasal grapheme.[5]

(3) Nasal vowel before labial and alveolar obstruents: [tsamba] /tsąba/ <tsąba> 'money'
   a. <tsamba>

(4) Nasal vowel before velar obstruents: [dɛŋges] /dęges/ <dęges> 'be quiet!'
   a. <dëges, dë̋gase, deges, dagasé>
   b. <dëgés, dëgéshǫ, nën nëgës, dęgés, dęges, dąges, dą́gës, daɻges>
   c. <dangase, dangásë>

Whenever the phoneme is realized phonetically as a nasal vowel, we again find considerable variability: nasal vowel (5a), nasal vowel + nasal consonant (5b), and oral vowel + nasal consonant (5c). Despite variability in both environments — (4) and (5) — transcribers' choices are not completely parallel: pattern (5b) (nasal vowel + nasal consonant) is not attested in the first environment; and pattern (4a) (oral vowel only) is not attested in the second environment. This suggests that transcribers' ad hoc decisions are not entirely random.[6]

(5) Nasal vowel: [hɻ(l)ðɪ] /hɻlðɪ/ <hɻldhɪ> 'you stop (doing it)!'
   a. <hɻltthë, hɻtthé, hɻthɪh, hɻdhɪ, hɻdhí>
   b. <hɻndhɪ, hɻ́nthɪ>
   c. <hɪndhɪ, hɪndhɪgh, hɪndhɪgh>

Overall, the initial transcripts thus indicate that transcribers are sensitive to allophonic variation: they adopt different transcription strategies in the case of nasal unpacking vs nasal vowels. Within each of these strategies, the principles of English orthography interfere: less familiar graphemes (nasal hooks, ë, tone) as well as missing graphemes (velar nasal) cause uncertainty. Choices are not random, though, and future studies should target the limits of variability. In particular, a

5. Similarly, Example (4) shows considerable variability with respect to the unfamiliar grapheme <ë> /ɛ/ [ɛ, ə] (written <ë, e, a, ë, ę, ą, aɻ>) and the high tone.

6. The spelling of the nasal consonant in (5) is likely to be correlated to the lateral segment /l/ before the stem, which has been lost in the pronunciation of many speakers, but leaves traces related to the realization of nasal vowels. The spelling of voiceless <th> (instead of voiced <dh>) probably arises because of a common paradigmatic alternation in other verbs (<dh> '2SG', <th> '1SG').

quantification of patterns would establish if some are more frequent than others (as was demonstrated for Iha and Qaqet).

An increasing familiarity with the Dëne Sųłiné orthography — accompanied by discussions with researchers — leads transcribers to abandon phonetic spellings in favour of the more standardized orthographic spelling. When comparing initial transcripts with later transcripts, we observe more and more instances of the hooked nasals even in cases of nasal unpacking. Similarly, we observe conventionalization of frequent forms; compare e.g. <sǫ́nt'ë> 'it is alright' (earlier spellings) with <sǫ́t'ë> (later spellings).

5. Discussion and conclusion

As a point of departure, this study asked to what extent transcription strategies provide a window into metalinguistic knowledge. Sapir had already proposed that speakers *hear* phonemes rather than phonetic elements; our aim was to observe whether speakers also *write* phonemes rather than phones when they transcribe. To this end, we focused on spelling variation in relation to nasalization in three unrelated languages: Iha, Qaqet, and Dëne Sųłiné. Analysing spelling variation is a way to use the transcript (as product) as evidence for what happens in transcription (as practice). And, if intuition about underlying phonemes is revealed through transcription, this study may provide empirical evidence for the commonly encountered assumption that speakers are aware of phonemes rather than phones. While the methodological variation between the three case studies means that the results cannot be directly compared, we are able to formulate hypotheses, and also to fulfil the ancillary goal of demonstrating variation in transcription procedures.

In Iha and Qaqet, we focused on prenasalized voiced plosives, which are phonemic in Iha, but phonetic realizations of voiced plosives in Qaqet. In Iha, underlying contrasts are partly obscured by considerable variability in the realization of phonemes: word-initially, both prenasalized plosives and oral voiced plosives can optionally be realized with or without prenasalization; word-internally, ambiguity arises between prenasalized plosives and sequences of nasal + homorganic (voiced or voiceless) plosive because the plosive part of prenasalized plosives may be realized either voiced or voiceless. This variability is not yet well understood. In Qaqet, allophonic variation is more restricted, as phonemic contrasts only collapse in one environment: voiced and voiceless plosives are both realized voiced following a nasal. This assimilation is very frequent word-internally, and may also occur across word boundaries. Taking the language-specific set of contrasting phonemes and allophonic variation into account, we investigated how

speakers transcribe plosives with regard to the spelling of (i) the nasal component and (ii) the voicing of the plosive. Specifically, we asked whether there are patterns in the distribution of alternative spellings and if we can detect an effect of Iha and Qaqet's different phonological systems on transcribers' choices.

In both languages, there was surprisingly little variability. In Iha, there were high levels of agreement between all three transcribers, whose spelling generally reflected the underlying phoneme. This was surprising for two reasons: first, because of the large number of allophones attested in the wider Iha corpus and, second, because Iha had a high rate of spelling differences between the three transcripts. Similarly, in Qaqet, spelling differences were highly frequent overall, but variability in the spelling of nasal components and voicing was more constrained. Hence, spelling differences in Iha and Qaqet must occur predominantly in other domains, which presents an avenue of future research. Having hypothesized that higher variability for nasal vowels in Dëne Sųłné was due to the unfamiliar hooked vowel graphemes, one such domain is variability in the use of (combinations of) graphemes absent in the orthography of the majority language, e.g. coarticulated stops in Iha (<qp>, <ngmb>) and the retroflex rhotic in Qaqet (<rl>).

In Iha, we investigated inter-transcriber variation. In general, deviation from the underlying oral plosive phonemes was due to reflecting voicing assimilation in the spelling. These were consistently written voiceless following consonants and voiced following vowels across all three transcripts, regardless of whether they followed a word boundary (i.e. a space) or not. The variation in the spelling of prenasalized plosives, however, was not consistent across transcripts: one transcriber had a clear preference for using voiceless plosives, while the other two consistently used the voiced plosive grapheme. Since voicing differences were so consistent between transcripts in other contexts, this systematic difference invites detailed analysis of the phonetic profile of the plosive portion in prenasalized stops in order to investigate whether it differs in some way from voiced plosives. It is furthermore notable that we did not find any differences due to dialectal background and/or familiarity with writing practice. One of the transcribers, PB, comes from a different dialect area and is least familiar with written Iha, but his strategies did not differ from those of the other two.[7]

7. We found such a difference in a domain not investigated in this contribution: PB systematically chose to write a different place of articulation for word-final plosives. This may reflect a dialectal difference, less familiarity with Iha writing (which led him to prioritize a more phonetic spelling), and/or the cross-linguistically attested greater importance of word-initial phonemes for word recognition (Hawkins & Cutler 1988). At present, it is not clear to what extent one or all of these factors contribute to the observed spelling variability in one domain (i.e. place of articulation in word-final plosives), but not in other domains (i.e. nasalization and voicing).

In Qaqet, we investigated intra-transcriber variation. In relation to prenasalization, transcribers strongly preferred to represent the phoneme (voiced plosive) rather than the phone (prenasalized plosive). Unlike in Iha, however, a good number of the respective alternative spellings were attested, too. For voiced plosives, writing the phonetic nasal component suggests that Qaqet speakers are highly aware of this phonetic feature. Conversely, not writing the nasal component in nasal + voiced plosive sequences may suggest that transcribers sometimes re-interpret sequences as singleton phonemes. This is possibly related to frequency effects (compared to voiced plosives, nasal + voiced plosive sequences are less frequent) and/or a low functional load (we are not aware of any minimal pair involving the opposition voiced plosive vs nasal + voiced plosive). In all cases, the alternative spelling thus does not affect word recognition. We furthermore investigated if Qaqet transcribers represent the voicing assimilation of voiceless plosives following nasals. Here, the results revealed the opposite tendency (similar to Iha): to transcribe the phonetic realization (voiced plosive) rather than the underlying phoneme (voiceless plosive). This tendency was much stronger within words compared to between words.

In Dëne Sųłné, our focus was on the spelling of nasal vowel phonemes. In the orthography, nasal vowels are rendered by hooked letters, regardless of their actual realization. Transcribers, by contrast, strongly preferred to represent allophonic variation in their writing. Whenever nasal vowels are phonetically realized as sequences of oral vowels + nasal consonants, transcribers predominantly wrote them as such in their transcripts, too. Interestingly, only the velar place of articulation patterned differently (represented variably as oral or nasal vowels, mostly without the nasal segment). When realized as nasal vowels, we find considerable spelling variability (nasal vowel, nasal vowel + nasal, oral vowel + nasal). We hypothesize that such variability is triggered by missing graphemes (in the case of the velar nasal) and unfamiliar graphemes (in the case of the nasal hook), as transcribers are forced to make ad hoc decisions on how to best represent their phonological and phonetic knowledge in writing.

The results of the three case studies paint an interesting picture in which there is no simple answer to the question of whether native speakers transcribe (in place of Sapir's "hear") phonemes or phones. The comparison between phonemic prenasalization in Iha and phonetic prenasalization in Qaqet showed that transcribers prefer to write phonemes: Iha transcripts almost always include the nasal portion; Qaqet transcripts rarely do. However, we also noticed an intriguing difference between the two languages in this respect: although Qaqet speakers predominantly write phonemes, too, the phonetic spelling emerges as a robust minority pattern. We may hypothesize that the phonemic oppositions within a language constrain inter- and intra-transcriber variability, i.e. we would expect the

variability to be greater where it does not affect word recognition (as in Qaqet), and to be less where a phonetic spelling may encroach into the domain of a different phoneme (as in Iha). The Qaqet result is also remarkable from a second perspective. The literature generally assumes that speakers are only aware of phonemic contrasts and/or allophones whose realization is very different from that of the underlying phoneme (see e.g. the contributions in Cahill & Rice 2014). In writing prenasalization, Qaqet speakers demonstrate awareness of a phonetic feature that is not contrastive and that accompanies the voiced plosive. This is unexpected and suggests that speakers may be aware of more phonetic detail than is often assumed.

With regard to voicing, there was evidence that transcribers prefer to write phones: in Qaqet, the voicing assimilation of voiceless plosives following nasals was frequently reflected in the transcript; in Iha, both voiced and voiceless plosives were transcribed with voicing assimilation at morpheme or word boundaries (voiceless following consonants, voiced following vowels). That is, we observe an exclusive or predominant phonemic spelling in one domain (prenasalization) vs a predominant phonetic spelling in another domain (voicing). It may be that voicing is more salient to speakers than prenasalization, thus facilitating a more phonetic spelling. However, this cannot be the complete answer, as Qaqet speakers demonstrate an awareness of phonetic prenasalization. Alternatively, it may again reflect language-specific phonologies, specifically regarding free variation of phones vs their complementary distribution. In both languages, the phonetic realization of prenasalization is somewhat less predictable than the realization of voicing. Transcription practices may reflect this predictability, with transcribers preferring to transcribe systematic phonological changes, but not unpredictable variation. The results from the Dëne Sųłiné case study support this hypothesis, as transcribers demonstrated sensitivity to systematic positional variation by rendering oral vowel + nasal consonant sequences (preceding obstruents) differently from nasal vowels (in other contexts). It is possible that a phonemic spelling would need to be actively learned in cases of complementary distribution; yet, speakers of other languages have been observed to reject predictable morphophonological processes being represented in the orthography (e.g. Snider 2014). The intriguing question remains of how to account for such differences, and whether it is possible to identify language-internal or external factors that make such automatic processes more or less salient to speakers.

The design of the Qaqet and Dëne Sųłiné studies allowed us to investigate intra-transcriber variation, which revealed differences in the development of transcription practices over time. Where variability remained the same between earlier and later Qaqet transcripts, it progressively lessened in the Dëne Sųłiné transcripts. We hypothesize that this is due to differences in the transcription

process arising from differences in language ideologies between the two communities. The Dëne Sųłiné transcribers received schooling within a standard language culture (Milroy 2001), in which it is widely accepted that words can be — and be spelt — "right" or "wrong". Consequently, transcribers actively sought feedback on their spelling, which the researcher provided. This feedback was incorporated into later transcripts and reinforced through further discussion with the researcher. There was no such feedback loop in the transcription process for Qaqet: the transcribers neither requested nor received feedback on their spelling. Theirs is not a standard language culture: there is only limited schooling in English, and writing is a rarity in the Qaqet transcribers' daily lives. These observations have implications for analyses based on transcripts created at different points in time, and in cultural contexts with differing attitudes towards language standardization.

This study opens several avenues for future research. Here, we concentrated on the spelling of nasalization, but the fact that a great deal of spelling variation came from other domains means that additional phenomena can be explored in this way. Furthermore, a considerable number of non-identical segments in the Qaqet and Iha datasets contained other kinds of differences, categorized as the addition, substitution, or repetition of material, and word division. In terms of the latter, an interesting observation from the Iha study was that voicing assimilation was quite consistently reflected in the transcripts, regardless of whether the word boundary was indicated by a space in the transcript or not. This differed from Qaqet, where there was a clear difference in the proportion of voicing assimilation following nasals reflected within words compared to across word boundaries.

While it is to be expected that at least some variation — both inter- and intra-transcriber — is purely accidental, we contend that systematic transcription variation has evidentiary value; that is, it can reveal metalinguistic knowledge. This variation may be between transcripts completed by the same native-speaking transcribers, as was the case in Qaqet and Dëne Sųłiné, or in transcripts by different speakers, as in Iha. But the variation may also be between a transcript completed in the standardized researcher orthography and a native-speaker transcript. All variation has both theoretical (for linguistic analysis) and practical (for orthography development or literacy programmes) implications. The decisions made by transcribers represent a "reflective mode" of language (Himmelmann 2018: 36) — and our exploratory case studies showed that transcripts provide a valuable source of insight into transcribers' metalinguistic knowledge, constituting a first step towards "reaching a better understanding of the transcription process itself and its relevance for linguistic theory" (Himmelmann 2018: 35).

Acknowledgements

First and foremost, we thank the members of the Iha, Qaqet, and Dëne Sųłiné communities who contributed to this research, either directly as transcribers or indirectly in any number of ways. We also thank two anonymous reviewers and our editor, Sonja Riesberg, for useful feedback that helped shape the final version of this chapter. Finally, we thank Nikolaus Himmelmann, both for his own research on transcription that inspired this contribution, and for the various ways in which he has provided support and guidance in our linguistic endeavours at every level — from MA thesis (KW), to best practices for sustainable field work (DJ), to professorship (BH).

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following abbreviations are used:

C+	consonant followed by morpheme boundary
NM	noun marker
POSD	possessed
POSS	possessor
RCD	reduced (noun class)
RPRT	reportative
SF	stem former
V+	vowel followed by morpheme boundary.

References

Blust, Robert. 2013. *The Austronesian Languages*. Canberra: Asia-Pacific Linguistics.

Bucholtz, Mary. 2000. "The Politics of Transcription." *Journal of Pragmatics* 32 (10): 1439–1465.

Bucholtz, Mary. 2007. "Variation in Transcription." *Discourse Studies* 9 (6): 784–808.

Cahill, Michael, and Keren D. Rice (eds.). 2014. *Developing Orthographies for Unwritten Languages*. Dallas: SIL International.

Chelliah, Shobhana L., and Willem de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Dordrecht: Springer.

Cook, Eung-Do. 2004. *A Grammar of Dëne Sųłiné (Chipewyan)*. Winnipeg, MB: University of Manitoba.

Crowley, Terry. 2007. *Field Linguistics: A Beginner's Guide*. Oxford: Oxford University Press.

De Vries, Lourens. 2015. "Views of Orality and the Translation of the Bible." *Translation Studies* 8 (2): 141–155.

Dobrin, Lise M. 2021. "The Arapesh Suitcase Miracle: The Interpretive Value of Reproducible Research." *Language Documentation and Description* 21: 37–69.

Evans, Nicholas, and Sebastian Fedden (eds.). To appear. *The Oxford Guide to the Papuan Languages*. Oxford: Oxford University Press.

Foley, William. 1986. *The Papuan Languages of New Guinea*. Cambridge: Cambridge University Press.

Haviland, John B. 1996. "Text from Talk in Tzotzil." In *Natural Histories of Discourse*, ed. by Michael Silverstein, and Greg Urban, 45–78. Chicago: University of Chicago Press.

Hawkins, John A., and Anne Cutler. 1988. "Psycholinguistic Factors in Morphological Asymmetry." In *Explaining Language Universals*, ed. by John A. Hawkins, 280–317. Oxford: Blackwell.

Hellwig, Birgit, Carmen Dawuda, Henrike Frye, and Steffen Reetz. 2023. *Qaqet Child Language Corpus: Longitudinal Study*. Cologne: Data Center for the Humanities.

Himmelmann, Nikolaus P. 2018. "Meeting the Transcription Challenge." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 33–40. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24806

Jung, Dagmar, and Nikolaus P. Himmelmann. 2011. "Retelling Data: Working on Transcription." In *Documenting Endangered Languages*, ed. by Geoffrey L. J. Haig, Nicole Nau, Stefan Schnell, and Claudia Wegener, 201–220. Berlin: De Gruyter Mouton.

Krauss, Michael E., and Jeff Leer. 1981. "Athabaskan, Eyak and Tlingit Sonorants." *Alaska Native Language Center Research Papers* 5.

Maas, Utz. 2009. "Orality Versus Literacy as a Dimension of Complexity." In *Language Complexity as an Evolving Variable*, ed. by Geoffrey Sampson, David Gil, and Peter Trudgill, 164–177. Oxford: Oxford University Press.

Milroy, James. 2001. "Language Ideologies and the Consequences of Standardization." *Journal of Sociolinguistics* 5 (4): 530–555.

Mosel, Ulrike. 2006. "Fieldwork and Community Language Work." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 67–86. New York: De Gruyter Mouton.

Mosel, Ulrike. 2015. "Putting Oral Narratives into Writing: Experiences from a Language Documentation Project in Bouganville, Papua New Guinea." In *Language Contact and Documentation*, ed. by Bernard Comrie, and Lucía Golluscio, 321–342. Berlin: De Gruyter Mouton.

Narfafan, Sutriani, and Emanuel Tuturop. 2009–2016. *DobeS Iha documentation*. The Language Archive, MPI Nijmegen. https://hdl.handle.net/1839/00-0000-0000-0017-EC35-B

Ochs, Elinor. 1979. "Transcription as Theory." In *Developmental Pragmatics*, ed. by Elinor Ochs, and Bambi B. Schieffelin, 43–72. New York: Academic Press.

Palmer, Bill (ed.). 2017. *The Languages and Linguistics of the New Guinea Area: A Comprehensive Guide*. Berlin: De Gruyter Mouton.

Paradis, Carole, and Jean-François Prunet. 2000. "Nasal Vowels as Two Segments: Evidence from Borrowings." *Language* 76 (2): 324–357.

Pawley, Andrew, and Harald Hammarström. 2017. "The Trans New Guinea Family." In *The Languages and Linguistics of the New Guinea Area: A Comprehensive Guide*, ed. by Bill Palmer, 21–196. Berlin: De Gruyter Mouton.

Pike, Kenneth L. 1947. *Phonemics: A Technique for Reducing Languages to Writing*. Ann Arbor: University of Michigan Press.

Reiman, D. Will. 2010. "Basic Oral Language Documentation." *Language Documentation and Conservation* 4: 254–268. http://hdl.handle.net/10125/4479

doi Rice, Keren D. 1993. "A Reexamination of the Feature [Sonorant]: The Status of 'Sonorant Obstruents'." *Language* 69 (2): 308–344.

Ross, Malcolm D. 1998. "Proto Oceanic: Phonology and Morphology." In *The Lexicon of Proto Oceanic: The Culture and Environment of Ancestral Oceanic Society*, Vol. 1, ed. by Malcolm D. Ross, Andrew Pawley, and Meredith Osmond, 15–36. Canberra: Pacific Linguistics.

doi Sapir, Edward. 1949. "The Psychological Reality of Phonemes." In *Selected Writings of Edward Sapir in Language, Culture, and Personality*, ed. by David G. Mandelbaum, 46–60. Berkeley: University of California Press.

doi Schultze-Berndt, Eva. 2006. "Linguistic Annotation." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 213–251. New York: De Gruyter Mouton.

doi Seifart, Frank. 2006. "Orthography Development." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 275–300. New York: De Gruyter Mouton.

Snider, Keith. 2014. "Orthography and Phonological Depth." In *Developing Orthographies for Unwritten Languages*, ed. by Michael Cahill, and Keren D. Rice, 27–48. Dallas: SIL International.

doi Stebbins, Tonya, Beth Evans, and Angela Terrill. 2017. "The Papuan Languages of Island Melanesia." In *The Languages and Linguistics of the New Guinea Area: A Comprehensive Guide*, ed. by Bill Palmer, 775–894. Berlin: De Gruyter Mouton.

doi Tabain, Marija, Marc Garellek, Birgit Hellwig, Adele Gregory, and Richard Beare. 2022. "Voicing in Qaqet: Prenasalization and Language Contact." *Journal of Phonetics* 91: 101138.

doi Tabain, Marija, and Birgit Hellwig. 2023. "Qaqet." *Journal of the International Phonetic Association* 53 (3): 1073–1094.

doi Usher, Timothy, and Antoinette Schapper. 2022. "The Greater West Bomberai Language Family." *Oceanic Linguistics* 61 (1): 469–527.

Walker, Katherine, and Nikolaus P. Himmelmann. To appear. "Iha." In *The Oxford Guide to the Papuan Languages*, ed. by Nicholas Evans, and Sebastian Fedden. Oxford: Oxford University Press.

Widodo, Supriyanto, Donatus Nimbitkendik, Philipus Kabes, Kaliktus Tanggahma, and Abas Bahamba. 2012. *Kamus Dwibahasa Bahasa Indonesia-Bahasa Iha* [*Bilingual Indonesian-Iha Dictionary*]. Jayapura: Balai Bahasa Provinsi Papua dan Provinsi Papua Barat.

PART III

Synchronic perspectives

Phonetics, phonology, and prosody

CHAPTER 15

Pauses, parts of speech, and word order

A comparative corpus study on 27 languages

Frank Seifart
CNRS-SeDyL | Humboldt University of Berlin

This study investigates how pauses before nouns and verbs interact with word order and prosodic structure across a worldwide sample of 27 languages from the DoReCo corpus. Whereas prior research has suggested that more pauses occur before nouns in general, the results here show that this asymmetry holds consistently only in verb-final languages, while in other languages, pausing patterns are mixed. Since word order mediates pausing asymmetries, it likely also plays a major role in explaining prefixing asymmetries — namely, the global dispreference for prefixes on nouns compared to verbs. The study demonstrates the potential of comparative corpus linguistics for examining typological patterns in speech, especially when incorporating prosodic annotation.

Keywords: pauses, parts of speech, word order, comparative corpus linguistics, affixation asymmetries

1. Introduction

Pauses — the silence or the *ehm*s etc. between words in speech — are in integral part of language. They create pragmatic meaning in conversation (Clark & Fox Tree 2002; Arnold et al. 2007), they build prosodic structure by marking intonation unit boundaries, along with other phonetic cues (Himmelmann & Ladd 2008; Himmelmann et al. 2018; Peck et al. 2021), and they interact closely with syntactic structure (Peck & Becker 2024). They also reflect increased processing and planning effort for upcoming linguistic material (Levelt et al. 1999). And they shape historical language change by inhibiting affixation processes which leads to typological asymmetries. This has been shown in Himmelmann's (2014) seminal study on prosodic motivations for the overall preference for suffixes, as opposed to prefixes, in the world's languages (Cysouw 2009; Hammarström 2021).

https://doi.org/10.1075/slcs.240.15sei
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

Himmelmann (2014) argues that the suffixing preference arises as a result of an asymmetry in prosodic phrasing: Prosodic breaks, in particular pauses, are more likely to occur between preposed function words and content words than between content words and postposed function words, compare *it happened* ***in (…) Cologne*** *(natural pause)* vs. *it happened* ***years (…)[2] ago*** (less natural pause). One reason for this difference is that pauses between proposed function words and content words are more effective floor-holding devices. Another is that content words have a higher affordance in language processing that induce hesitations, compared to function words. The higher probability of prosodic breaks and pausing in the former context prevents diachronic processes of prefixation of (formerly) free function words compared to potential suffixation in the latter context. For instance, it prevents *in* from becoming a case prefix. Accordingly, suffixation is synchronically much more common among the world's languages compared to prefixation, although this asymmetry is more pronounced for some categories, like case markers, compared to others, like person markers (Cysouw 2009).

Taking up the idea that differences in the probability of pauses are intimately linked to the part of speech of the following word, Seifart et al. (2018) have shown that pauses before nouns tend to be more frequent than before verbs, and have linked this observation to a higher processing demand of nouns compared to verbs, specifically to the function of nouns to introduce new information, compared to verbs, which may represent new or old information. This suggests that a similar motivation as that proposed by Himmelmann (2014) for the suffixing preferences may be involved in explaining the dispreference of prefixes on nouns, compared to verbs. This dispreference had been documented in world-wide languages samples by Cutler et al. (1985: 730) and in Balthasar Bickel's work reported in Seifart et al. (2017).

Earlier work on affixing asymmetries by Bybee et al. (1990), which, however, did not take pausing into account, had indicated that basic word order plays an important role in the complex process of historical affixation processes. Namely, they found that "V-medial languages fail to affix preposed grams [to verbs] at the same rate as V-final and V-initial languages" (Bybee et al. 1990: 7). Himmelmann (2014: 933) also linked this pattern to prosodic phrasing. He suggests that prosodic breaks at clause boundaries, specifically pauses preceding verbs in verb initial languages, facilitate the prosodic phrasing of preceding function words with the following lexical hosts. In this case it would thus not be hesitation pauses that are responsible for inhibiting affixation, but prosodic breaks — often involving, presumably, pauses — at intonation unit boundaries that in fact facilitate affixation. Even though Bybee et al.'s (1990) data were limited to verbs only, and the precise role of pauses in this context has not been explored in detail, these observations

indicate the importance of taking word order, as well as intonation units, into account in the study of pausing patterns connected to parts of speech.

The current study thus investigates how word order interacts with pause probabilities before the two major word classes, nouns and verbs, and with their position within intonation units. It also takes into account other factors known to influence pausing such as word length and word frequency. Results of this may eventually contribute to clarifying aspects of the extremely complex question of historical affixation processes and the resulting affixing asymmetries, although the current study does not directly address affixation.

Bringing word order into the picture necessitates significantly widening the cross-linguistic scope compared to previous studies. The current study investigates pause probabilities before nouns vs. verbs in a sample of 27 languages from the DoReCo database (Seifart et al. 2022), including languages from five macro-areas, a total of 18 linguistic families, and including SOV, SVO, and VSO languages, as well as languages with no dominant word order. It applies a comparative corpus approach (Schnell & Schiborr 2022), and is based on data compiled within the framework of Documentary Linguistics (Himmelmann 1998), most of them from small and endangered languages. These are multimedia corpora documenting more or less spontaneously produced traditional spoken genres. These corpora were created, usually through field work, and annotated and are authored by language experts. They have been subsequently processed to make them compatible and usable for comparative analyses. Crucial steps in this process included semi-automatic time-alignment of transcription with audio at the level of phones, morphs, and words (within DoReCo), and — for the purpose of the current study — the integration of automatic annotation of Intonation Unit breaks produced by Inbar et al. (2025), which facilitates to study non intonation-unit initial hesitation pauses separately.

The following sections introduces the data and their annotation in more detail (Section 2), followed by a presentation of the raw distribution of pauses before nouns and verbs in the data (Section 3), then a statistical analysis (Section 4), a discussion (Section 5), and a conclusion (Section 6).

2. Data and annotation

The input data for the current study (Table 1, Figure 1) consist of spoken corpora of, on average, about 10,000 words for 27 language taken from the DoReCo database (Seifart et al. 2022). These corpora comprise mostly narrative texts that were time-aligned at the phone level, including pause identification, through automatic procedures with subsequent manual checking (Paschen et al. 2020). The 27 datasets used in the current study were selected from the set of 51 DoReCo datasets using the following two criteria: (i) They include annotation of part of speech by language experts (which is the case for 31 of these languages); and (ii) Intonation Unit breaks were annotated for them by Inbar et al. (2025) (which is the case for 46 of these languages). The inclusion of these annotations also explains why the 1.2 version of DoReCo was used rather than the more recent 2.0 version (Seifart et al. 2024). Inbar et al. (2025) excluded multiple-speaker recordings from their dataset to simplify their acoustic analysis, and so did I. The great majority of data used here are either traditional or personal narratives, but also included are some texts that had been elicited with the help of stimuli such as the Pear Story film.

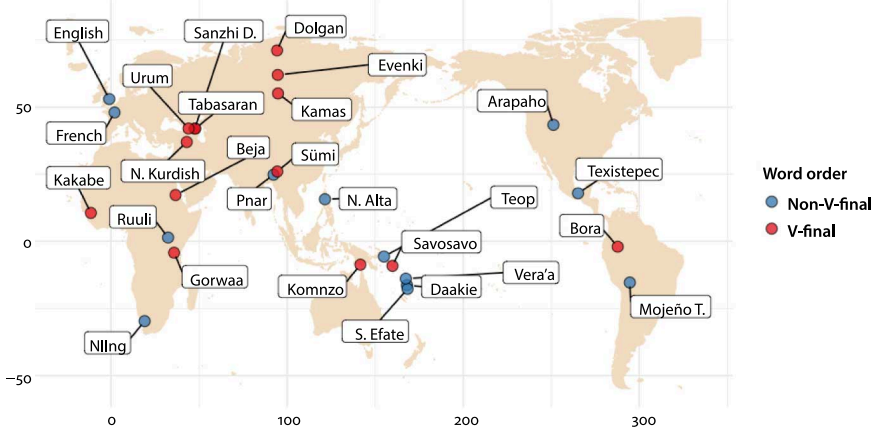


Figure 1. Location and basic word order of languages used in the current study

I identified the basic word order for each language (see Table 1) based on Dryer (2013) for the languages included there, and on published sources on the languages where it was not, applying Dryer's (2013) criteria. (One of them, Bora, has equally frequent SOV and OSV orders, but is treated as SOV here.) It would have been preferrable, in line with the token-based approach taken here, to use word order preferences computed from the corpus data themselves. However, it is

Table 1. Information on the languages and corpora used in the current study, sorted by area*

	Language	Language information			Corpus information	
		BWO	Glotto-code	Family	Reference	Words
1	Beja	SOV	beja1238	Afro-Asiatic	Vanhove 2022	15,466
2	Gorwaa	SOV	goro1270	Afro-Asiatic	Harvey 2022	19,992
3	Ruuli	SVO	ruul1235	Atlantic-Congo	Witzlack-Makarevich et al. 2022	8,746
4	Kakabe	SOV	kaka1265	Mande	Vydrina 2022	46,635
5	N‖ng	SVO	nngg1234	Tuu	Güldemann et al. 2022	27,049
6	Pnar	SVO	pnar1238	Austroasiatic	Ring 2022	20,527
7	Northern Kurdish (Kurmanji)	SOV	nort2641	Indo-European	Haig et al. 2022	9,658
8	English (Southern England)	SVO	sout3282	Indo-European	Schiborr 2022	35,296
9	French (Swiss)	SVO	stan1290	Indo-European	Avanzi et al. 2022	13,339
10	Sanzhi Dargwa	SOV	sanz1248	Nakh-Daghestanian	Forker & Schiborr 2022	15,191
11	Tabasaran	SOV	taba1259	Nakh-Daghestanian	Bogomolova et al. 2022	5,059
12	Sümi	SOV	sumi1235	Sino-Tibetan	Teo 2022	14,452
13	Evenki	SOV	even1259	Tungusic	Kazakevich & Klyachko 2022	8,377
14	Dolgan	SOV	dolg1241	Turkic	Däbritz et al. 2022	18,876
15	Urum	SOV	urum1249	Turkic	Skopeteas et al. 2022	18,805
16	Kamas	SOV	kama1351	Uralic	Gusev et al. 2022	39,853
17	Northern Alta	NoDom	nort2875	Austronesian	Garcia-Laguia 2022	11,141
18	Daakie	SVO	port1286	Austronesian	Krifka 2022	12,039
19	Nafsan (South Efate)	SVO	sout2856	Austronesian	Thieberger 2022	27,200
20	Teop	SVO	teop1238	Austronesian	Mosel 2022	12,149
21	Vera'a	SVO	vera1241	Austronesian	Schnell 2022	17,782
22	Savosavo	SOV	savo1255	Isolate	Wegener 2022	12,081
23	Komnzo	SOV	komn1238	Yam	Döhler 2022	33,762
24	Mojeño Trinitario	SVO	trin1278	Arawakan	Rose 2022	16,557
25	Bora	SOV	bora1263	Boran	Seifart 2022	29,366
26	Texistepec Popoluca	VSO	texi1237	Mixe-Zoque	Wichmann 2022	8,468
27	Arapaho	NoDom	arap1274	Algic	Cowell 2022	10,430

* From top to bottom, the areas are: Africa, Eurasia, Papunesia, South America, North America. Glottocodes are language identification codes provided by the Glottolog (Hammarström et al. 2024), from where information on family and area is also taken. In the column on basic word order (BWO), NoDom stands for no dominant word order. The column Words refers to the number of word form tokens.

currently not possible to obtain these because the data used here do not contain annotation of clause boundaries. What has been done is to determine the position of nouns and verbs with respect to IU boundaries (see Appendix).

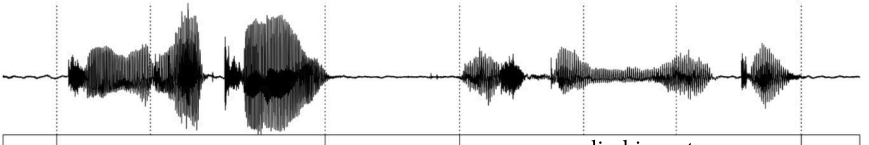
I grouped the 27 languages into 14 verb-final ones and 13 non-verb final languages (10 SVO, one VSO, and two with no dominant word order). The distribution of word order in this sample matches closely the world-wide distribution of word orders: Hammarström's (2016) analysis, based on 5,252 languages, but stratified by language families, yielded 56.8% verb-final languages (56.6% SOV + 0.2% OSV) vs. 43.4% non verb-final ones (13.0% SVO + 6.3% VSO + 3.5% VOS + 0.7% OVS + 6.1% no dominant order). Regarding verb final vs. non verb-final languages my sample is thus fairly representative (51.9% vs. 48.1%), although within non-verb final languages, SVO is notably overrepresented (37.0%) while VSO (3.7%) and VOS (none) languages are underrepresented.

I then transformed the manual annotations of parts of speech that were provided by the creators of the corpora into four general labels for the purpose of the current study: N(oun), V(erb), PRO(noun), and OTHER. For instance, in the Kakabe dataset (Vydrina 2022), some of the language-specific part of speech tags from Vydrina's (2022) annotation that were categorized by me as N(ouns) were *n.loc, n.prop*, and *n -mrph -mrph*, and some of the ones categorized as V(erbs) were *v, vi, vt, vi -mrph*, and *cop*, while *itj, quot, adv*, and *adj* were all categorized as OTHER. In the final analysis, only N(oun) and V(erb) tokens were retained. Data were further processed and analyzed in R (R Core Team 2024). In this process, I removed items labelled as foreign material, sung, false starts, etc. during manual correction within DoReCo (Paschen et al. 2020: 2660).

Next, I added annotations of intonation unit boundaries provided by Inbar et al. (2025). They adopted and adapted an existing algorithm for automatic annotation of intonation units (Suni et al. 2017). Their method is based on the automatic detection of pitch, loudness, and word duration and a mathematical tool called wavelet transformation which assigns a boundary strength score to each word boundary. The distribution of scores was found to be bimodal, and this distribution was used to classify each word boundary in the data as being strong, i.e. as corresponding to an intonation unit boundary, vs. weak, i.e. as an intonation-unit internal word boundary. Inbar et al. (2025) tested the accuracy of the method by comparing its performance on spontaneous speech from four languages with the results of manual expert annotations of the same data. The method reached accuracies of 84.24% for English (USA), 73.00% for Hebrew, 77.55% for Russian, and 94.25% for Totoli. These figures are indicative of the reliability of this annotation, especially the figure for Totoli, which were fieldwork data similar to those of the DoReCo languages used here, even though none of these four test cases were among the DoReCo datasets used here (Totoli was only included in a later

DoReCo release, Bardají i Farré et al. 2024). While pauses appear to have played an important role in the IU detection using this method, crucially for the current study, there are mismatches between pauses and IU boundaries (see Appendix).

The identification of pauses in the current study directly follows annotations already present in DoReCo data. In DoReCo, instructions for manual coders that corrected the first round of automatic time alignment included to make sure that the start and end time of all perceivable silent pauses were correctly marked. During manual correction, filled pauses were also manually annotated, i.e. non-lexical hesitation markers like English *ehm* and *uhm* (Clark & Fox Tree 2002). Both types of pauses were included in the current study. I excluded pauses that were longer than two seconds, to exclude silent intervals that are likely not part of coherent speech, and pauses that were shorter than 50 milliseconds, to exclude "articulatory pauses" (Hieke et al. 1983) that might have been overlooked during manual annotation. Figure 2 illustrates the annotation of data used in the current study.



	one man			was climbing a tree			
	timi	lakhi		asbo	mgho	ake	
	n	cardnum		n	v	post	
	N	OTHER		N	V	OTHER	
	person	one		tree	climb	NMLZ	
<p:>	IU initial		<p:>	IU initial			<p:>

Figure 2. Annotated data from Sümi (Teo 2022: ABT3-IA1_pearstory01-A), illustrating time-alignment of word start and end times, language-specific part of speech annotation and their transformation to N, V, and OTHER labels, as well as annotation of silent pauses (<p:>) and intonation unit-(IU-) initial position. Figure created with Praat (Boersma & Weenink 2025)

3. Raw distribution of pauses before nouns and verbs

Before carrying out statistical analyses, I inspected pause occurrences in the data. Regarding the distribution of intonation unit breaks and pauses in the data used here, it is noteworthy that in almost all languages, most intonation unit boundaries (before nouns and verbs), as identified by Inbar et al. (2025), coincided with pauses, with considerable cross-linguistic variation, ranging from 49% in French

to 98% in Northern Alta (see Appendix). These differences are clearly related to the overall number of pauses in the data, which is low for French and high for Northern Alta, as apparent from Figure 3, as well as from the ratio of IU boundaries per pauses, also given in the Appendix. Another noteworthy fact reported in the Appendix is that most pauses in the data used here co-occur with intonation unit boundaries, leaving a minority of pauses in intonation-unit internal positions, again with some cross-linguistic variation, ranging from 4% intonation-unit internal pauses in English to 29% in Dolgan (again only counting pauses before nouns and verbs). Explaining these differences is outside the scope of the current study, but note here that they reflect high variability in the data.

Figure 3 gives on overview of the proportions of pauses before nouns vs. verbs found in the data used here. This figure shows, firstly, great variability across languages regarding pausing. For instance, overall pauses probabilities range from less than 10% (for both nouns and verbs) in French to between more than 35% before nouns to almost 50% before verbs in Arapaho. These differences are likely partially due to word length, Arapaho being a polysynthetic language. Regarding differences between pauses before nouns vs. verbs, these figures already indicate the importance of word order. According to these figures, in all SOV languages in the sample (Beja, Bora, Dolgan, Evenki, Gorwaa, Kakabe, Kamas, Komnzo, Nothern Kurdish, Sanzhi Dargwa, Savosavo, Sümi, Tabasaran, and Urum), there are more pauses before nouns than before verbs. In the remaining languages, the pausing patterns are diverse: In the two languages with no dominant word order (Arapaho and Northern Alta), there are more pauses before verbs compared to nouns, just as in five out of the ten SVO languages in the sample (French, Mojeño Trinitario, Pnar, Ruuli, and Teop). The remaning five SVO languages (Daakie, English [from Southern England], N‖ng, South Efate, and Vera'a) display the expected pattern (more pauses before nouns), just as the only VSO language Texistepec Popoluca. Note, however, that in many languages the differences in the raw data are too small to draw any conclusions, which is one reason to carry out statistical analyses. Another, even more important one, is that in SOV languages, nouns are more likely to occur intonation-unit initially than verbs (see Appendix), which is where most pauses occur. Therefore, we cannot tell from the raw data whether the difference between SOV and other languages shown in Figure 3 reflects simply intonation-unit initiality or an effect of part of speech.

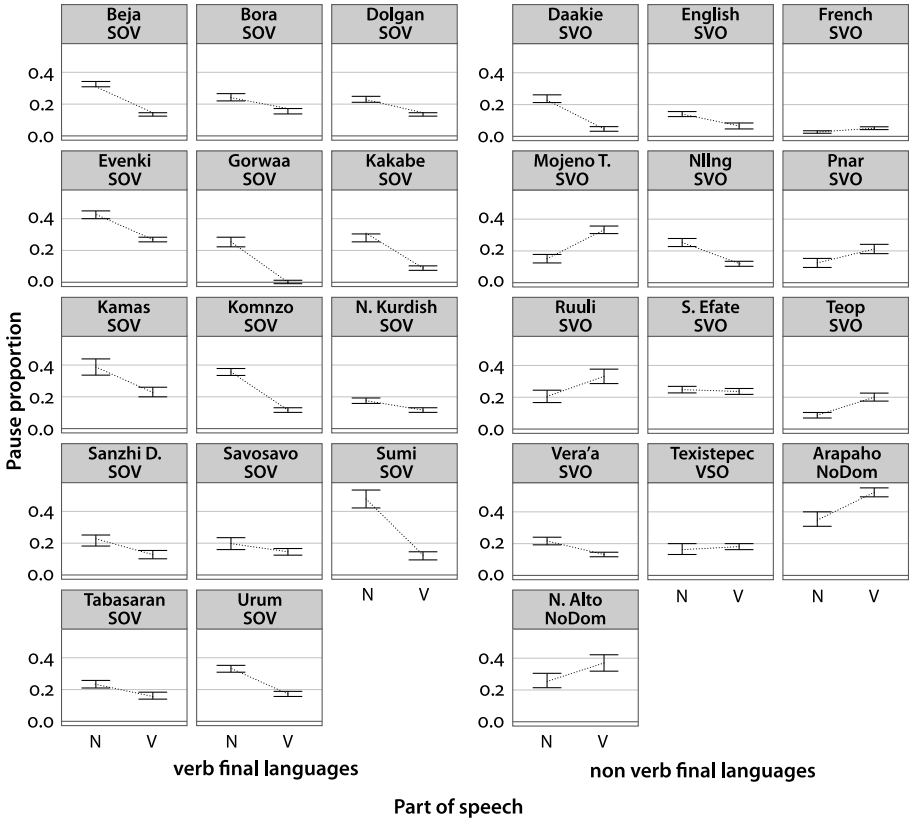


Figure 3. Proportion of nouns (N) and verbs (V) with a preceding pause, with languages grouped by word order: Verb-final (left) and non-verb-final (right). Dotted lines connect mean values. Error bars represent bootstrapped 95% confidence intervals that were calculated by resampling the data 1,000 times and taking the 2.5th and 97.5th percentiles of the resulting distribution of means

4. Statistical analyses

To analyze pause probabilities and their relation to parts of speech, word order, and other factors, I fitted a Bayesian regression model with the structure given in (1) using the R library brms (Bürkner 2017). Priors were set to reflect expectations for a strong positive effect for intonation-unit initial position, a weak negative effect for word length, a weak positive effect for word frequency, as well as default priors for other fixed effects and a standard prior for random effect standard deviations.

(1) pause before ~ part of speech * verb final
+ part of speech * intonation-unit initial
+ word length
+ word frequency
+ (1 + part of speech | language)
+ (1 | speaker)

The factors included in this model were defined as follows:

a. *Part of speech* is coded as verb vs. noun, as described above. I am interested here in whether and how the effect of part of speech interacts with two other factors, the basic word order of a languages, and the target word's position within an intonation phrase. Therefore, part of speech is included in the model as part of interactions with these two factors.
b. Basic word order was included as a binary factor, *verb final* vs. non-verb final, which was coded as described above.
c. *Intonation-unit initial* vs. non intonation-unit initial positions were defined as described above.
d. Longer words (more syllables or phonemes) require more extensive phonological planning, increasing the likelihood of hesitation pauses before articulation (Levelt 1989: 259). Therefore, *word length* is included in the statistical model. It is calculated as the number or phones per orthographic word. Phones are obtained from DoReCo's conversion of orthographic transcriptions to X-SAMPA symbols for the purpose of automatic time-alignment. Word boundaries follow experts' transcription. Both in terms of phonological representation (using orthographic characters) and in terms of word boundaries, the language documentation corpora used here generally follow standards of modern descriptive linguistics. For English and French, the MAUS automatic aligner used by DoReCo produced phonological representations based on pronunciation dictionaries that were used here to calculate word length in number of phones. For better statical modelling, this factor was centered.
e. Similar to word length, *word frequency* is expected to influence pause probabilities in that less frequent words are more likely to be preceded by hesitation pauses. Frequency is used here instead of contextual predictability, which is arguably more directly related to hesitation pauses (Beattie & Butterworth 1979), but requires more textual data than what is available for the majority of the languages studied here. Word frequency was calculated as the relative frequency of word forms in the entire text material available in DoReCo for each language, including texts from DoReCo's "extended" set, i.e. texts that are not

time-aligned (Table 1), and also texts that were excluded by Inbar et al. (2025) because they included more than one speaker. This dataset included a total of over half a million word tokens, with on average close to 19,000 words per language.

f. *Language* is included as a group-level ("random") slope so that the effect of part of speech on the probability of a pause occurring before a word can vary across languages. This accounts for the expectation that in some language there may be overall more pauses than in others, as well as the expectation that there may be differences in the effect that part of speech (noun vs. verb) has on pausing.

g. A final factor accounts for random differences between individual *speakers*. The way this factor is operationalized reflects that I expect that some speakers will pause more often than others (for a variety of possible reasons, including age), but I do not expect that for some speakers the effect of part of speech on pausing will be different.

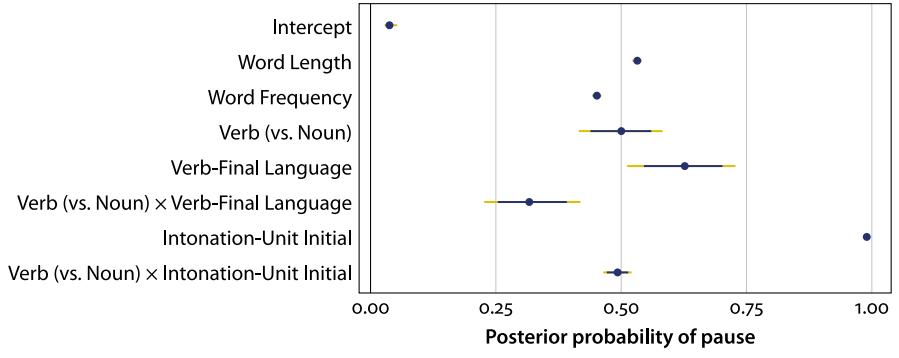


Figure 4. Results of Bayesian regression model for the effect of various factors on the posterior probability of pauses, with 95% (yellow) and 85% (blue) HPDI intervals

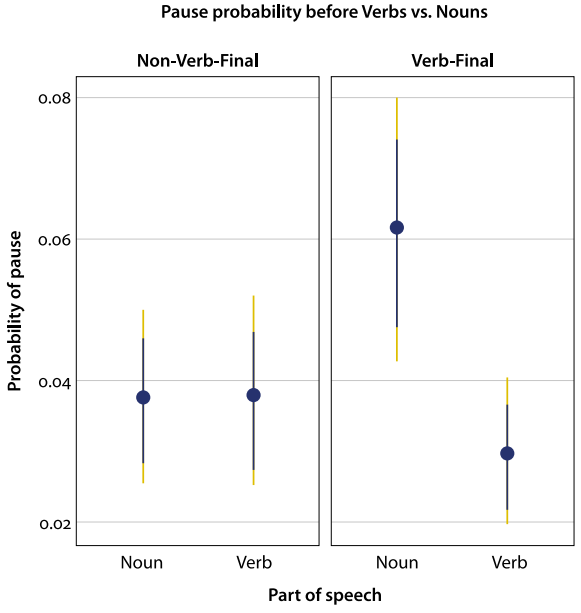
Results obtained from the statistical model given in (1) are summarized in Figure 4. Starting from the top, this Figure gives a result for the intercept, which corresponds to a noun of average word length and frequency, in a non-verb final language, and in a non intonation-unit initial position. Next, it shows that longer words induce more preceding pauses and more frequent words induce fewer pauses. The pausing patters predicted by these two factors are clearly hesitation pauses, not pauses employed for intonation-unit boundary marking. It is reassuring that the analyses captures these hesitation pauses because there are much fewer of these, compared to intonation-unit boundary-marking pauses, as can be seen from the high correlation between pauses and intonation unit boundaries in the data (see Appendix).

We now move on to the factors of main theoretical interest for the current analysis. First, it can be observed that part of speech by itself, across the entire data set, does not have a credible effect on pause probabilities: The mean effect is close to zero, and both 95% and 85% Highest Posterior Density Intervals (HPDIs) overlap with zero. HPDIs represent the most credible ranges of the estimated parameter values, given the model and the data. Following conventions in linguistic research (e.g. Donnelly et al. 2024), effects are interpreted as showing weak evidence if the 85% HPDI excludes zero, and strong evidence if the 95% HPDI does. Next, we see that in verb-final languages, there are overall more pauses preceding nouns and verbs. A likely reason for that will be discussed further below. The result for the interaction of being a verb with being a verb-final language indicates a clear effect of word order on the differential probability of pauses: In verb-final languages, there are fewer pauses preceding verbs than pauses preceding nouns, as we will also discuss in more detail below.

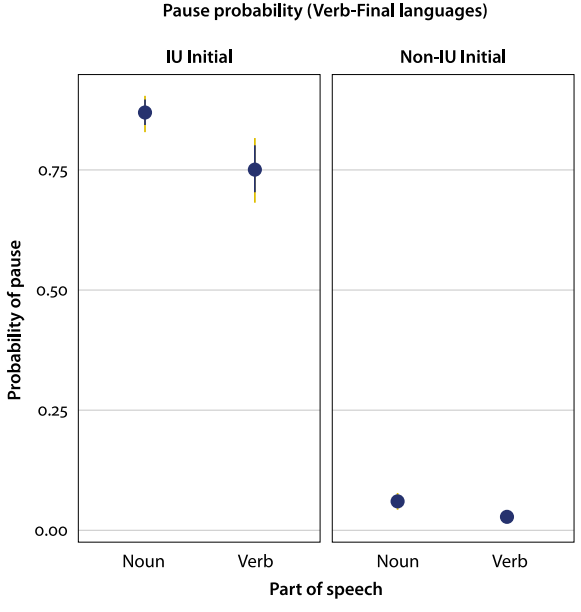
The position of a word in an intonation unit has, as expected, a strong effect on the probability of pauses, stronger than any other factor. Given that between 49% and 98% of intonation unit boundaries coincide with pauses in the data (see Section 2 and Appendix), it is not surprising that being in an intonation-unit initial position raises the probability of pauses drastically for both nouns and verbs. Interestingly, however, as the bottom row in Figure 4 shows, there is no evidence that the position in an intonation unit would affect the difference in pausing before nouns vs. verbs (both 85% and 95% HPDIs include zero).

For a deeper understanding of the two interactions, we now break down the results by word order and by position relative to intonation unit boundaries, respectively, in Figure 5. Figure 5a shows that in non-verb-final languages, we found no credible difference in pause probability between verbs and nouns, with a negligible mean difference of 0.0016 and almost entirely overlapping 95% and 85% HPDI intervals. This suggests that pause probability is not influenced by part of speech across the group of non-verb-final languages. In contrast, in verb-final languages, nouns do have a credibly higher pause probability than verbs, with a mean difference of –0.0328 and non-overlapping 95% and 85% HPDI intervals. These results again suggest that the relationship between pause probability and part of speech varies depending on the verb-finality. Figure 5a also shows that pauses before verbs in verb-final languages are within the same range of pauses before both nouns and verbs in non verb-final languages, but that there are many more pauses before nouns in verb-final languages. This explains why there are more pauses preceding nouns and verbs, taken together, in verb-final languages, as the results in Figure 4 showed.

We now need to take a closer look at the role of intonation units, because, after all, in verb final languages, verbs are less likely to be intonation-unit initial and



a.



b.

Figure 5. Pause probabilities before nouns vs. verbs in different word orders (a), and — within verb-final languages — in different positions relative to intonation

unit (IU) boundaries (b), indicating mean probabilities and 95% (yellow) and 85% (blue) HPDI intervals

nouns are more likely to be intonation-unit initial, which is where most pauses occur. In Figure 2, an example of this from Sümi is given, a verb-final language. In this example, a verb is preceded by its arguments within the same intonation unit, which is preceded by a pause. The question is thus whether the difference observed in Figure 5a is merely due to the position of nouns and verbs in intonation units, which in turn is heavily influenced by word order, rather than by part or speech itself.

To investigate this, we break results further down by looking at verb-final languages only, and, among these, whether the pausing difference between nouns and verbs persists in intonation-unit initial positions and/or non intonation-unit initial positions. Figure 5b shows, firstly, that both nouns and verbs have a much higher probability to be preceded by a pause in intonation-unit initial positions, as expected from then high correlation between IU boundaries and pauses in the data (see Appendix). Figure 5b also shows that in both intonation-unit initial positions and non intonation-unit initial positions there are differences in pausing between nouns and verbs. In SOV languages, verbs may be less likely to occur in intonation-unit initial positions, but when they do, they are less likely to be preceded by pauses compared to nouns in this position. The mean difference in this position is −0.123, i.e. there are on average 12% more pauses before nouns than before verbs, with non-overlapping 95% HPDI intervals. In non-intonation-unit initial positions, pauses are overall much rarer, as expected, and it may be that nouns are less likely to occur in these positions in SOV languages. But in these positions, too, there are credibly more pauses before nouns than before verbs. The mean difference in pause probability was −0.033, i.e. 3% more pauses before nouns compared to verbs, with non-overlapping 85% and 95% credible intervals.

In contrast to this, additional testing, not shown in Figure 5, showed that in non verb-final languages, pause probabilities before nouns are almost identical to those before verbs, both in intonation-unit initial positions (mean difference −0.0047) and intonation-unit medially (mean difference 0.0002), with overlapping 85% HPDIs in both cases.

5. Discussion

The expectation that there are more pauses before nouns than before verbs was clearly born out in verb-final languages — which is the most common word order across the world's languages. But it was not borne out across the group of languages with other word orders, including 10 SVO languages, one VSO language

and two with no dominant word order. The languages within this group — as well as the SVO languages alone — are mixed in terms of noun-verb pausing patterns. As a result of this, in SVO languages, and non verb-final languages generally, there is no credible difference between pause probabilities before nouns compared to verbs. And because of this, across the entire sample, no credible difference was found between pauses before nouns compared to verbs. These results are in line with Seifart et al.'s (2018) findings, although that study used a slightly different methodology: On the one hand, they looked at pauses in 500 ms windows preceding nouns and verbs, which might include pauses preceding short (function) words that precede the target word. On the other hand, they looked only at pauses that were internal to so called annotation units. The precise definition of these units varies across the corpora, but they generally consist of stretches of speech that are intonationally, grammatically, and pragmatically complete and may comprise an entire turn. Notably, annotation units very often include more than one intonation unit. Therefore, their results likely included both non intonation-unit initial and intonation-unit initial pauses. Across the nine languages in that study, pauses were significantly more likely to occur before nouns in all four SOV languages, Even [even1260], Hoocąk [hoch1243], Bora (also included in the current study), and Chintang [chhi1245]. Among the three SVO languages, there were more pauses before nouns than verbs in N‖ng, but not in Dutch, where there was no significant difference, and English (US), which showed a reversed effect. In the VOS language Baure [baur1253], there was no significant difference and in the VSO language Texistepec (also included in the current study) there were again more pauses before nouns compared to verbs.

Regarding the interpretation of these results, the current study, based on the larger language sample used here, was able to show that noun-verb pausing asymmetries do not necessarily hold across languages with different word orders. This indicates that the explanation for this pausing asymmetry involves additional factors that interact with Seifart et al.'s (2018) interpretation for this asymmetry. They suggested that paradigmatic complexity and in particular referential choice and pronominalization in nouns, but not in verbs, is responsible for more pauses before nouns. On the other hand, the current study's findings are consistent with Seifart et al.'s (2018) argumentation that word complexity alone do not explain increase in pausing, as earlier studies had argued in connection with the expectation that there should be more pauses before verbs compared to nouns because of verbs' greater complexity: The statistical analyses applied here showed that part-of-speech, in interaction with word order, has an effect in addition to word length.

The current study additionally showed that the effect of word order on noun-verb pausing differences is not just a side effect of intonation-unit initiality, at least not according to the intonation-unit boundary annotation used here: In verb-

final languages, there were more pauses before nouns than before verbs not only intonation-unit initially, but the difference was also found in non intonation-unit initial positions. This strongly suggests that the difference — at least in SOV languages — is due to a difference linked to part of speech, i.e. noun vs. verb.

6. Conclusion

Word order has a considerable influence on the pausing patterns associated with the two major parts of speech, nouns and verbs: In verb-final languages, noun use clearly elicits more pauses than verb use, but not necessarily in other languages. There may thus be properties of nouns, compared to verbs, that lead to increased pausing, but these effects may be overridden by effects linked to word order. Since verb final is the most common word order among of the world's languages, these findings presented here may play a role in explaining the global prefixing dispreferences on nouns, compared to verbs, along the lines of Himmelmann's (2014) prosody-based motivation for the suffixing preference. But the explanation as a whole is likely going to be complex. This explanation will also have to account for potential differences in how pauses at intonation unit boundaries affect affixation, compared to the intonation-unit internal hesitation pauses that Himmelmann (2014) focuses on. These two types of pauses are often viewed as fundamentally distinct. However, results from the current study indicate that, for pausing patterns before nouns and verbs, intonation-unit initial pauses (whose function is presumably prosodic-boundary marking) behave very similarly to intonation-unit internal pauses (presumably signs of hesitation). This suggests that the occurrence of these two types of pauses might be sensitive a common underlying common factor at some level. Further progress in the study of these pauses might necessitate refined annotation of intonation-unit boundaries than the ones available in the current study, which involve considerable error margins.

Another noteworthy finding here is the internal consistency regarding noun-verb pausing patterns among verb-final languages vs. the mixed behavior among other languages. For the time being, it remains an open question why non-verb final languages behave so inconsistently — and overall differently — compared to verb-final languages regarding pausing asymmetries. However, note that this situation also parallels affixation patterns: "The group of V-final languages in our sample is much more consistent in its positioning and affixation tendencies than the V-medial group" (Bybee et al. 1990: 7). This suggests that full explanations for the role of pauses in affixation processes might need to take into account other, yet unidentified characteristics that differentiate noun-pausing languages from verb-

pausing languages within the diverse set of V-medial, and maybe non verb-final languages more generally.

The current study aimed to push the extension of documentary linguistics towards comparative corpus linguistics one step further by enriching fieldwork-based spoken corpora with advanced (semi-)automatic annotation techniques. This allows for the study of spoken language variability across a representative sample of the world's languages. This approach not only highlights the importance of small languages for our understanding of language and languages but also honors the efforts of linguist who contributed to documenting them.

Funding

The current study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project numbers SE 1949 3-1 and SE 1949 5-1 —, and the ANR-DFG program "projets franco-allemand en sciences humaines et sociales / deutsch-französisches Programm in den Geistes- und Sozialwissenschaften" — project number ANR-18-FRAL-0010-01.

Acknowledgements

I am grateful to Jan Strunk and Balthasar Bickel for extensive discussions leading up to the current study, as well as to Sarah Wesolek and Frederic Blum for help with the statistics. Errors are all mine.

Data availability

All code and custom annotation used in the current study are available at https://doi.org/10.5281/zenodo.17764707. Inbar et al.'s (2025) intonation-unit annotation can be found at https://osf.io/g2v4f in the folder Inbar_et_al_2025/Data_osf/IUs_wd/. DoReCo 1.2 files can be downloaded at https://nakala.fr/10.34847/nkl.7cbfq779.v10 (please make sure to select the correct version [1.2] for each dataset).

References

Arnold, Jennifer E., L. Hudson, and Michael K. Tanenhaus. 2007. "If You Say thee uh You Are Describing Something Hard: The On-line Attribution of Disfluency During Reference Comprehension." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33 (5): 914–930.

Avanzi, Mathieu, Marie-José Béguelin, Gilles Corminboeuf, Federica Diémoz, and Laure Anne Johnsen. 2022. "French (Swiss) DoReCo Dataset." In Seifart et al. (2022).

Bardají i Farré, Maria, Christoph Bracks, Claudia Leto, Datra Hasan, Sonja Riesberg, Winarno S. Alamudi, and Nikolaus P. Himmelmann. 2024. "Totoli DoReCo Dataset." In *Language Documentation Reference Corpus (DoReCo) 2.0*, ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Beattie, Geoffrey W., and B. L. Butterworth. 1979. "Contextual Probability and Word Frequency as Determinants of Pauses and Errors in Spontaneous Speech." *Language and Speech* 22 (3): 201–211.

Boersma, Paul, and David Weenink. 2025. *Praat: Doing Phonetics by Computer* [Computer program]. Version 6.4.27. http://www.praat.org/. (28 March 2025).

Bogomolova, Natalia, Dmitry Ganenkov, and Nils Norman Schiborr. 2022. "Tabasaran DoReCo Dataset." In Seifart et al. (2022).

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28.

Bybee, Joan L., William Pagliuca, and Revere D. Perkins. 1990. "On the Asymmetries in the Affixation of Grammatical Material." In *Studies in Typology and Diachrony: Papers Presented to Joseph H. Greenberg on his 75th birthday*, ed. by William A. Croft, Suzanne Kemmer, and Keith Denning, 1–42. Amsterdam: John Benjamins.

Clark, Herbert H., and Jean E. Fox Tree. 2002. "Using uh and um in Spontaneous Speaking." *Cognition* 84 (1): 73–111.

Cowell, Andrew. 2022. "Arapaho DoReCo Dataset." In Seifart et al. (2022).

Cutler, Anne, John A. Hawkins, and Gary Gilligan. 1985. "The Suffixing Preference: A Processing Explanation." *Linguistics* 23 (5): 723–758.

Cysouw, Michael. 2009. "The Asymmetry of Affixation." *Snippets* 20 (3): 10–14.

Däbritz, Chris Lasse, Nina Kudryakova, Eugénie Stapert, and Alexandre Arkhipov. 2022. "Dolgan DoReCo Dataset." In Seifart et al. (2022).

Döhler, Christian. 2022. "Komnzo DoReCo Dataset". In *Language Documentation Reference Corpus (DoReCo) 1.2*, ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Berlin/Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Donnelly, Seamus, Caroline Rowland, Franklin Chang, and Evan Kidd. 2024. "A Comprehensive Examination of Prediction-Based Error as a Mechanism for Syntactic Development: Evidence from Syntactic Priming." *Cognitive Science* 48 (4): e13431.

Dryer, Matthew S. 2013. "Order of Subject, Object and Verb." In *The World Atlas of Language Structures Online*, ed. by Matthew S. Dryer, and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/81. (30 December 2022).

Forker, Diana, and Nils Norman Schiborr. 2022. "Sanzhi Dargwa DoReCo Dataset." In Seifart et al. (2022).

Garcia-Laguia, Alexandro. 2022. "Northern Alta DoReCo Dataset." In Seifart et al. (2022).

Güldemann, Tom, Martina Ernszt, Sven Siegmund, and Alena Witzlack-Makarevich. 2022. „N‖ng DoReCo Dataset." In Seifart et al. (2022).

Gusev, Valentin, Tiina Klooster, Beáta Wagner-Nagy, and Alexandre Arkhipov. 2022. "Kamas DoReCo Dataset." In Seifart et al. (2022).

Haig, Geoff, Maria Vollmer, and Hanna Thiele. 2022. "Northern Kurdish (Kurmanji) DoReCo Dataset." In Seifart et al. (2022).

Hammarström, Harald. 2016. "Linguistic Diversity and Language Evolution." *Journal of Language Evolution* 1 (1): 19–29.

Hammarström, Harald. 2021. "Measuring Prefixation and Suffixation in the Languages of the World." In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, ed. by Ekaterina Vylomova, Elizabeth Salesky, Sabrina Mielke, Gabriella Lapesa, Ritesh Kumar, Harald Hammarström, Ivan Vulić, Anna Korhonen, Roi Reichart, Edoardo Maria Ponti, and Ryan Cotterell, 81–89. Online: Association for Computational Linguistics.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. *Glottolog 5.1*. Leipzig.

Harvey, Andrew. 2022. "Gorwaa DoReCo Dataset." In Seifart et al. (2022).

Hieke, Adolf E., Sabine Kowal, and Daniel C. O'Connell. 1983. "The Trouble with 'Articulatory' Pauses." *Language and Speech* 26 (3): 203–214.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2014. "Asymmetries in the Prosodic Phrasing of Function Words: Another Look at the Suffixing Preference." *Language* 90 (4): 927–960.

Himmelmann, Nikolaus P., and D. Robert Ladd. 2008. "Prosodic Description: An Introduction for Fieldworkers." *Language Documentation & Conservation* 2 (2): 244–274. http://hdl.handle.net/10125/4345

Himmelmann, Nikolaus P., Meytal Sandler, Jan Strunk, and Volker Unterladstetter. 2018. "On the Universality of Intonational Phrases: A Cross-linguistic Interrater Study." *Phonology* 35 (2): 207–245.

Inbar, Maya, Eitan Grossman & Ayelet N. Landau. 2025. "A Universal of Speech Timing: Intonation Units Form Low-Frequency Rhythms." *Proceedings of the National Academy of Sciences* 122 (34): e2425166122.

Kazakevich, Olga, and Elena Klyachko. 2022. "Evenki DoReCo Dataset." In Seifart et al. (2022).

Krifka, Manfred. 2022. "Daakie DoReCo Dataset." In Seifart et al. (2022).

Levelt, Willem J.M. 1989. *Speaking: From Intention to Articulation*. Cambridge, Mass., and London: MIT Press.

Levelt, Willem J.M., Ardi Roelofs, and Antje S. Meyer. 1999. "A Theory of Lexical Access in Speech Production." *Behavioral and Brain Sciences* 22 (1): 1–38.

Mosel, Ulrike. 2022. "Teop DoReCo Dataset." In Seifart et al. (2022).

Paschen, Ludger, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. "Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo)." In *Proceedings of the 12th Language Resources and Evaluation Conference*, ed. By Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2657–2666. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.324

Peck, Naomi, and Laura Becker. 2024. "Syntactic Pausing? Re-Examining the Associations." *Linguistics Vanguard* 10 (1): 223–237.

Peck, Naomi, Kirsten Culhane, and Maria Vollmer. 2021. "Comparing Cues: A Mixed Methods Study of Intonation Unit Boundaries in Three Typologically Diverse Languages." Presentation, AG 10a Prosodic Boundary Phenomena at the 43rd Annual Conference of the German Linguistic Society, Freiburg, 23–26 February 2021.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Ring, Hiram. 2022. "Pnar DoReCo Dataset." In Seifart et al. (2022).

Rose, Françoise. 2022. "Mojeño Trinitario DoReCo Dataset." In Seifart et al. (2022).

Schiborr, Nils Norman. 2022. "English (Southern England) DoReCo Dataset." In Seifart et al. (2022).

Schnell, Stefan. 2022. "Vera'a DoReCo Dataset." In Seifart et al. (2022).

Schnell, Stefan, and Nils Norman Schiborr. 2022. "Crosslinguistic Corpus Studies in Linguistic Typology." *Annual Review of Linguistics* 8: 171–191.

Seifart, Frank. 2022. "Bora DoReCo Dataset." In Seifart et al. (2022).

Seifart, Frank, Ludger Paschen, and Matthew Stave (eds.). 2022. *Language Documentation Reference Corpus (DoReCo) 1.2*. Berlin/Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Seifart, Frank, Ludger Paschen, and Matthew Stave (eds.). 2024. *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Seifart, Frank, Jan Strunk, and Balthasar Bickel. 2017. "Recurrent Patterns in the Distribution of Speech Pauses Cause Languages to Develop more Prefixes in Verbs than in Nouns." Presentation, 12th Meeting of the Association for Linguistic Typology, Canberra, Australia, 14 December 2017.

Seifart, Frank, Jan Strunk, Swintha Danielsen, Iren Hartmann, Brigitte Pakendorf, Søren Wichmann, Alena Witzlack-Makarevich, Nivja H. de Jong, and Balthasar Bickel. 2018. "Nouns Slow Down Speech cross Structurally and Culturally Diverse Languages." *Proceedings of the National Academy of Sciences of the United States of America* 115 (22): 5720–5725.

Skopeteas, Stavros, Violeta Moisidi, Nutsa Tsetereli, Johanna Lorenz, and Stefanie Schröter. 2022. "Urum DoReCo Dataset." In Seifart et al. (2022).

Suni, Antti, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. "Hierarchical Representation and Estimation of Prosody using Continuous Wavelet Transform." *Computer Speech & Language* 45: 123–136.

Teo, Amos. 2022. "Sümi DoReCo Dataset." In Seifart et al. (2022).

Thieberger, Nick. 2022. "Nafsan (South Efate) DoReCo Dataset." In Seifart et al. (2022).

Vanhove, Martine. 2022. "Beja DoReCo Dataset." In Seifart et al. (2022).

Vydrina, Alexandra. 2022. "Kakabe DoReCo Dataset." In Seifart et al. (2022).

Wegener, Claudia. 2022. "Savosavo DoReCo Dataset." In Seifart et al. (2022).

Wichmann, Søren. 2022. „Texistepec Popoluca DoReCo Dataset." In Seifart et al. (2022).

Witzlack-Makarevich, Alena, Saudah Namyalo, Anatol Kiriggwajjo, and Zarina Molochieva. 2022. "Ruuli DoReCo Dataset." In Seifart et al. (2022).

Appendix

Overview of data retained in current study*

Language	BWO	Nouns		Verbs		IU boundaries		Pauses		Ratio IU boundaries/ pauses
		IU initial		IU initial		With pauses		IU internal		
Beja	SOV	4,164	36%	6,083	17%	2,485	70%	2,170	19%	1.15
Bora	SOV	1,333	23%	1,866	18%	622	77%	613	22%	1.01
Dolgan	SOV	2,170	24%	2,653	14%	863	67%	817	29%	1.06
Evenki	SOV	2,085	43%	2,718	25%	1,468	82%	1,570	23%	0.94
Gorwaa	SOV	713	26%	522	1%	183	95%	186	6%	0.98
Kakabe	SOV	1,564	26%	2,184	10%	589	92%	622	13%	0.95
Kamas	SOV	359	39%	684	20%	265	86%	297	23%	0.89
Komnzo	SOV	2,228	36%	2,039	12%	993	97%	1,035	7%	0.96
N. Kurdish	SOV	2,224	20%	2,230	16%	740	78%	629	8%	1.18
Sanzhi D.	SOV	593	24%	854	14%	226	88%	227	13%	1.00
Savosavo	SOV	502	21%	1,118	20%	318	76%	263	8%	1.21
Sümi	SOV	333	47%	713	20%	294	69%	242	16%	1.21
Tabasaran	SOV	1,186	21%	1,234	16%	439	91%	451	11%	0.97
Urum	SOV	2,657	26%	2,960	15%	1,109	89%	1,269	22%	0.87
N‖ng	SVO	1,084	29%	2,152	14%	590	80%	525	10%	1.12
Pnar	SVO	576	18%	817	23%	284	74%	240	12%	1.18
Daakie	SVO	1,453	24%	2,239	6%	455	87%	431	9%	1.06
S. Efate	SVO	1,986	28%	2,980	25%	1,203	90%	1,175	7%	1.02
English	SVO	2,732	23%	1,335	12%	758	58%	456	4%	1.66
French	SVO	1,719	5%	1,843	7%	195	49%	118	19%	1.65
Teop	SVO	1,079	9%	1,307	22%	376	86%	355	9%	1.06
Vera'a	SVO	1,401	25%	2,185	17%	683	79%	583	7%	1.17
Ruuli	SVO	470	16%	371	30%	180	96%	214	19%	0.84
Mojeño T.	SVO	965	13%	1,610	30%	610	91%	670	17%	0.91
Texistepec	VSO	423	16%	1,694	20%	386	81%	371	16%	1.04
Arapaho	NoDom	420	32%	1,562	48%	843	85%	949	25%	0.89
N. Alta	NoDom	408	21%	370	34%	213	98%	238	13%	0.89

* Including information on basic word order (BWO), numbers of nouns and verbs and the percentage of their occurrence in intonation unit (IU) initial position, as well as the number of IU boundaries and pauses (both before nouns and verbs only), sorted by basic word order. For corpus references, see Table 1 in the main text.

CHAPTER 16

Intonation contour length as a quantification of macro-rhythm

Constantijn Kaland
University of Cologne

Macro-rhythm, the tonal rhythmicity within phrases, constitutes a typological dimension in prosody. However, little empirical research supports this concept. Work on the rhythm-class hypothesis (i.e., stress-, syllable and mora-timing) and on macro-rhythm quantification based on regular intervals (isochrony) has led to unsatisfactory results. This study therefore abandons the assumption that macro-rhythmic differences should be sought in some form of temporal regularity. The current exploratory approach to macro-rhythm quantification uses a measure of *f0 contour length* on five languages (Greek, German, European Portuguese, French and Seoul Korean). This measure generally distinguishes the languages in the expected way, and its potential and limitations are discussed in the context of the prosodic typology of rhythm and its applicability to under-documented languages.

Keywords: macro-rhythm, prosodic typology, isochrony, f0 contour, contour length increase, Greek, German, European Portuguese, French, Seoul Korean

1. Introduction

Languages differ with regard to their prosody. The literature offers different ways to typologize prosodic differences between languages. An influential model of prosodic typology by Jun (2005) separated lexical (i.e., word) from post-lexical (i.e., phrase) prosody. The distinction between word- and phrase-level prosody is widely adopted in the literature (autosegmental-metrical models; e.g., Silverman et al. 1992) and has been quantified using acoustic correlates for specific phenomena across languages. For example, word-level stress is often captured by duration and intensity(-based) metrics (e.g., Gordon & Roettger 2017), whereas lexical tone and (post-)lexical pitch accents are mostly described by their fundamental frequency (f0) (e.g., Ladd 2008, Gordon 2014, Van Heuven 2018). More recently,

https://doi.org/10.1075/slcs.240.16kal
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

the typological model of prosody by Jun (2014) added another dimension — *macro-rhythm* — in order to capture differences between languages that have a similar word- and phrase-level prosody but nevertheless have a different tonal rhythm (Table 1). Very few attempts have been done to quantify macro-rhythm using acoustic measures, covering only three languages so far: English, Italian and Spanish (Prechtel 2020b; Polyanskaya et al. 2020). Crucially, the existing macro-rhythm quantifications are grounded in the idea that the speech signal exhibits a form of regularity that can be captured by acoustic measures. While this idea does find support in work on speech rhythm in general, it has not been sufficient to typologize languages in the way proposed in Jun (2014). In particular, the role of durational variability (i.e., *timing*) might have been overestimated in research on speech rhythm (Arvaniti 2009: 59; Rathcke & Smith 2015). The typological model that included macro-rhythm (Jun 2014) could benefit from alternative approaches to quantify tonal rhythmicity. The current paper explores a quantification of macro-rhythm without timing metrics, i.e., by taking into account the length of an intonation contour. To this end, five languages with supposedly varying degrees of macro-rhythm (Greek, German, European Portuguese, French and Seoul Korean) are compared using both scripted and unscripted speech. The following introductory sections summarize the main findings from the speech rhythm literature (1.1), discuss the concept of macro-rhythm (1.2), its quantification attempts (1.3) and outlines an alternative approach to (macro-)rhythm (1.4). Section 2 reports the methodology used in this study, Section 3 lists the results and the final Section 4 provides a discussion and conclusion.

1.1 Temporal and tonal approaches to rhythm

Speech rhythm, and in particular its acoustic manifestation, has been a challenge to understand. Assuming that rhythm could be measured by stable timed intervals between syllables, morae or stresses (i.e., isochrony; Pike 1945), research has failed to accurately classify languages, in particular in the search for rhythm metrics. It is beyond the scope of this chapter to review the details of these attempts. The reader is referred to excellent overviews by e.g., Arvaniti (2009) and Fletcher (2010). At least two conclusions can be drawn from speech rhythm research that are relevant to the current study. They both relate to the role of (absolute) duration as measured in the speech signal and are further summarized here.

Research has shown that regularity as measured by interval durations in the speech signal adds little to the way rhythmic differences between languages can be captured. Thus, there is no more isochrony between stressed syllables in so called stress-timed languages than in so called syllable-timed languages (e.g., Dauer 1983). Since then, many more attempts have been done to find metrics that could

distinguish these two alleged rhythm classes, in particular taking into account timing differences *within* syllable structure (e.g., Ramus et al. 2000; Low et al. 2000; Dellwo et al. 2007, Arvaniti 2012). It has been questioned, however, whether measures of temporal regularity alone are able to capture speech rhythm, in particular when assuming the above-mentioned rhythm classes, and whether rhythm should rather be approached as a perceptual phenomenon (e.g., Arvaniti 2009). This view is particularly supported by studies showing that listeners can perceive isochrony even with some degree of temporal *ir*regularity, as they tend to perceive short intervals as longer than they are and long intervals as shorter than they are (e.g., Woodrow 1951).

In this respect, an important theoretical distinction has been made; i.e., the separation of timing and rhythm (Arvaniti 2009). While timing refers to durational characteristics, rhythm is the regularity that can be derived from these characteristics. Arvaniti (2009) suggests that, although timing is an important part in rhythm research, it should be disentangled from rhythm by abandoning the traditional rhythm classes (i.e., stress- vs. syllable- vs. mora-timed languages). It neatly follows from the separation of timing and rhythm to include other acoustic correlates than duration into the quantification of rhythm (e.g., Kohler 2009). The literature indeed moved towards the study of tonal aspects of rhythm. For example, an important finding by Cumming (2011) showed that durational and tonal cues are interdependent in the perception of rhythm and that their relative strength differs per language. This was shown by a series of experiments in which Swiss German and Swiss French participants indicated the naturalness of a sentence's rhythm, without having received any explanation of the concept of rhythm. Acoustic manipulations were carried out in which syllable duration and f0 peak height was shortened or lengthened, both together and separately. These manipulated (deviant) versions were compared to an unmanipulated (non-deviant) version of the same sentence. Results showed that the highest naturalness scores were obtained when both duration and f0 were non-deviant. The degree of tolerance for deviant stimuli was different between the languages in that for Swiss German participants durational deviance worsened the naturalness scores more than tonal deviance, whereas for Swiss French participants both cues had similar effects, in particular when shortened/lowered. This result was interpreted as a genuine difference in the way rhythm is perceived across languages, i.e., not as a result of different acoustic manipulations in the stimulus materials of each language.

Similar conclusions were drawn in a production-perception study comparing Bulgarian, English and German (Barry et al. 2009). This study tested the sensitivity of a metric for durational variability of adjacent syllables (PVI, see details in Section 1.3) to differences in metrical structure (temporal) and accent patterns (tonal) in read and reiterated prose. Although the PVI metric appeared to be sen-

sitive to both, it was non-specific for the tonal component. That is, a sentence with de-accentuation in the first part and stronger accentuation in the second part would result in the same PVI measure as that sentence without either accent modification. A follow-up perception experiment with participants of the same three languages tested the relative contribution of several acoustic cues to perceived rhythm. This was done in a task in which participants compared two sequences of eight syllables and indicated which had a stronger rhythm. The sequences had a neutral (non-rhythmic) and a trochaic version, which were obtained by acoustic manipulation of syllable duration, f0 height, intensity level and vowel quality. Results showed that for English and German duration was a stronger rhythm cue than the others, whereas for Bulgarian duration and f0 were equally important cues. Intensity and vowel quality were weak cues in all languages tested.

Taken together, speech rhythm research has opened up to the inclusion of multiple acoustic cues, showing intricate differences between languages. The outcomes challenge the attempts to typologize languages based on rhythm and the quantification using rhythm metrics in particular. The next sections discuss an influential model of prosodic typology in which tonal aspects of rhythm are a key typological dimension (*macro-rhythm*; Jun 2014) and discuss the studies that quantified it.

1.2 Macro-rhythm

Essentially, three aspects make up the model of prosodic typology in Jun (2014): prominence type, word prosody and macro-rhythm (Table 1). The first two refer to the use of prosodic prominence at the phrase-level and the word-level respectively. Prominence type refers to the places of prosodic prominence; either the head, the edge, or both. Word prosody separates whether a language has stress, tone/lexical (pitch) accent, both or none of these. For example, Greek (Arvaniti & Baltazani 2005), German (Grice et al. 2005), and European Portuguese (Vigario & Frota 2003) are all classified as head languages with word stress, as they mark the phrasal head and make use of word stress. Cantonese is classified as a head language which makes use of lexical tone (Wong et al. 2005). All head languages analyzed in Jun (2014) make use of either stress, tone or both. French is an example of a language that only makes use of phrase prosody, marking its head and/or edge (Jun & Fougeron 2002). An example of an edge language is Korean (Kim & Cho 2009). Edge languages do not make use of word prosody as they only mark the edges of phrases (i.e., no (post-)lexical heads). Using just prominence type and word prosody as distinguishing features, the typological model "could not capture the fact that Greek has more regular phrase-medial tonal patterns than English" (Jun 2014: 521). In a similar vein, differences in tonal regularity between the head/edge languages French (high regularity) and Chickasaw (low regularity) could

not be accounted for. Furthermore, Spanish and Bengali are equally rhythmic phrase-medially, although they make use of different prominence types (head, and head/edge respectively). For these reasons, a third feature *macro-rhythm* was introduced to account for these differences and similarities.

Macro-rhythm refers to speech rhythm created by phrase-medial tonal regularity (Jun 2014: 522) and finds support in empirical studies showing that perceived speech rhythm depends on f0 patterns (e.g., Barry et al. 2009; Cumming 2011). Jun (2014) lists three factors that determine the degree of macro-rhythm, all depending on the way a language uses pitch accents, accentual phrase (AP) tones and/or word tones (henceforth "accents"). These three factors are the *number, type*, and *frequency* or *domain* of accents. As for the number of accents, languages are considered to be more macro-rhythmic when their inventory of accents is small, because that results in overall less variation in phrase-medial f0 contours. As for the type of accents, the most commonly used one will determine the degree of macro-rhythm; with rising or falling accents resulting in more rhythmicity than level accents. As for the frequency or domain of accents, languages in which every word receives an accent are considered to be more macro-rhythmic than those that apply accents in a larger domain, e.g., the accentual phrase containing multiple words.

The degree of macro-rhythm varies the most among head languages. For example, Greek is considered to be more macro-rhythmic than German because its most common accent is bi-tonal and involves a rising component (L*+H), whereas in German the most common accent is a single target on one (high) level (H*). German, in turn, is considered to be more macro-rhythmic than European Portuguese because the former generally has a smaller domain of accents (a content word or larger) than the latter (larger than a content word). Note that the frequency/domain factor is not decisive alone, but determines the degree of macro-rhythm in relation to the other factors. Among head/edge languages, for example, the accentual domain in French is also a content word or larger (as in German), although it is considered strongly macro-rhythmic, i.e., more than Chickasaw (medium macro-rhythm). The difference between these languages is mainly the result of a fixed accent position in French and a variable one in Chickasaw.

It is furthermore important to note that the three factors determining the degree of macro-rhythm do not all contribute with equal strength. This point is not discussed explicitly in Jun (2014), but follows from the examples discussed there. The issue pertains primarily to the number of accents (i.e., pitch accent inventory size). For example, Catalan is reported to have a relatively large inventory of pitch accents (six in total) whilst still being classified as a having a strong degree of macro-rhythm. Other head-languages in this category have smaller inventories, with Egyptian Arabic and Samoan as extremes (single accent inventory). What makes Catalan a strong macro-rhythmic language is the shape of its most commonly used pitch accent (rise with delayed peak) as well as the rela-

tively high frequency/small domain of accents (a content word or larger). In addition, for all prominence types and for each degree of macro-rhythm there are languages with an inventory size of five accents: Italian, Georgian, Oirat Mongolian (strong), English (varieties), Dutch, Bininj Gun-wok (medium), European Portuguese (weak). This again shows that it is easy to find languages for which accent inventory size is only a weak determiner of macro-rhythm. Thus, in the context of three factors influencing the degree of macro-rhythm, it follows that the shape and frequency of accents in a phrase play a more determining role for macro-rhythm than the composition of the inventory from which they are taken. Therefore, the influence of the first factor discussed above (number of accents) on the degree of macro-rhythm can be overruled by the type of most frequent accent. Number of accents, therefore, only holds as a general trend; with larger inventories there is a higher chance of more accent variability in the contour. The role of accent inventory on macro-rhythm is further discussed in Section 1.5.

Table 1. Prosodic typology in Jun (2014: 535). Shaded cells indicate language categories that are unexpected to be attested

Prominence type	**Word prosody**	**Macro-rhythm**		
		Strong	**Medium**	**Weak**
Head	Stress	Brazilian Portuguese, Castilian Spanish, Catalan, Egyptian Arabic, Greek, Italian, Samoan	Dutch, English, German, Jamaican Creole English, Lebanese Arabic	European Portuguese, Wolof, Kuot
	Tone / lexical pitch accent	Kihehe, Safwa, Somali	Isthmus Zapotec, Kimatuumbi, Skou	Cantonese
	Both	Papiamentu, Swedish		Mandarin
	None			
Head/edge	Stress	Bengali, Georgian, Kiche, Tamil, Tongan	Bininj Gun-wok, Dalabon	
	Tone / lexical pitch accent	Japanese, Leketio Basque		
	Both	Serbo-Croatian	Chickasaw	
	None	French, Standard Basque		
Edge	None	Accentless dialects of Japanese, Halh Mongolian, Oirat Mongolian, Seoul Korean, West Greenlandic		

1.3 Quantification of macro-rhythm

Jun (2014) proposed two ways of quantifying macro-rhythm. One is based on standard deviations (σ) of the intervals between fo peaks (p) and between f0 valleys (v) as well as the σ's of the rising (r) and falling (f) f0 slopes. The sum of all these σ's was called Mac*R*_Var and is expected to produce a higher value for languages with a weaker degree of macro-rhythm (i.e., larger σ's overall), see Equation 1. The second metric was called Mac*R*_Freq and captures the average number of high/low (HL) alternations per prosodic word (PW) within an intonational phrase (IP). This is calculated by dividing the number of f0 peaks in an IP by the number of PWs in an IP. In this way, Mac*R*_Freq values closer to 1 indicate strong macro-rhythmicity, see Equation 2. To date, two studies are known that made use of (one of) the macro-rhythm metrics proposed in Jun (2014) and/or other rhythm metrics to investigate macro-rhythmic differences between languages. Both studies compared English; either to Italian (Polyanskaya et al. 2020) or to Spanish (Prechtel 2019; Prechtel 2020b; Prechtel 2020a). These studies are discussed in the following, as well as a small number of other studies investigating the notion of macro-rhythm in different ways.

(1) Mac*R*_Var = $\sigma_p+\sigma_v+\sigma_r+\sigma_f$

(2) Mac*R*_Freq = $\frac{N\left(p\right)_{IP}}{N\left(PW\right)_{IP}}$

In Polyanskaya et al. (2020) unscripted retellings of an animated cartoon videoclips were analyzed (10 participants per language). The retelling lasted approximately 50 seconds per participant. The intonation contours produced in recorded retellings were stylized and the f0 turning points (targets) were automatically detected using the MOMEL algorithm (Hirst 2007). Several metrics relating to either the temporal or tonal domain were applied to the f0 targets to investigate whether these would reflect the supposedly higher degree of macro-rhythm in Italian compared to English (Table 2). Apart from Mac*R*_Var, also normalized pairwise variability index (*nPVI*; Low et al. 2000; Grabe & Low 2002) was calculated for the timing of H and L targets, both separately and taken together. The nPVI (Equation 3) captures pairwise variability between consecutive measurements with m as the number of measurements and d as the value of the kth measurement. The difference between two consecutive measurements is divided by the mean of these measurements to normalize speech rate and account for speaker differences. In addition, the variation coefficient (*Varco*) was calculated for the timing of H and L targets, both separately and taken together. Varco is the standard deviation of the measurements divided by their mean (Equation 4).

In the tonal domain, the f0 excursion width was measured as the difference in f0 between two consecutive targets and the number of f0 turning points was calculated whilst controlling for phrase length.

$$(3) \quad \text{nPVI} = 100 * \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{\left(d_k - d_{k+1}/2 \right)} \right| / (m-1) \right]$$

$$(4) \quad \text{Varco} = \frac{\sigma_d}{\mu_d}$$

Table 2. Macro-rhythm metrics applied to different domains (temporal/tonal/both), their application to the measured tonal targets (f0 turning points), and their reported effects, i.e., statistically significant difference between Italian and English in Polyanskaya et al. (2020) and between Spanish and English in Prechtel (2020b). Prechtel (2019) and Prechtel (2020a) are older versions reporting the same results and not listed here. Metrics not applied/reported are indicated with a hyphen '–'

Domain	Metric	Targets	Italian-English	Spanish-English
Temporal	NPVI	H, L, all	* (L, all)	n.s.
	Varco	H, L, all	n.s.	–
Tonal	F0 excursion width	Consecutive	*	–
	N targets per IP	All	*	–
	Mac*R*_Freq	All	–	*
Both	Mac*R*_Var	All	n.s.	n.s.
	↳ slope	H	–	n.s
	↳ slope	L	–	n.s
	↳ p-p intervals	H	–	*
	↳ v-v intervals	L	–	*

Results (Table 2) showed that in Italian, f0 targets are distributed more regularly in the temporal domain than in English, with significant differences found only for nPVI applied to L targets and all targets. F0 excursion width and the number of targets per IP were also found to be significantly higher in Italian than in English. Mac*R*_Var did not reveal significant differences between the languages. The latter outcome was attributed to the combination of temporal and tonal variability in this metric, although they are separate correlates of rhythm (e.g., Cumming 2011). It was furthermore suggested that Italian and English have similar sized tonal inventories, which could have resulted in comparable macro-rhythmicity degrees (i.e., the first factor defining macro-rhythm in Section 1.2).

In Prechtel (2020b), nPVI, Mac*R*_Var and Mac*R*_freq measures were compared using English and Spanish radio news recordings of one speaker per language (Table 2). F0 turning points were annotated manually following the procedure reported in Mennen et al. (2012). The nPVI metric did not show significant differences between the languages, which was attributed to the lack of f0 plateau measures in this metric (i.e., only turning points). The Mac*R*_Var metric showed a marginally significant difference in that higher variability was found for Spanish than for English, which was counter to the expectation (Table 1). The individual measures of the H and L slopes and intervals that are needed for the calculation of Mac*R*_Var (1) showed that Spanish had significantly shorter target intervals than English. Mac*R*_Freq confirmed a higher degree of macro-rhythmicity due to significantly more peaks per PW in Spanish than in English.

Another study did not quantify macro-rhythm using acoustic metrics, but rather compared the relationship between macro-rhythm and focus marking strategies in four typologically diverse languages: American English, Paraguayan Guaraní, Moroccan Arabic, and K'iche' (Burdin et al. 2015). The ones with stronger macro-rhythm appeared to mark focus using cues that did not disturb the regularity of f0 H and L alternations, i.e., deaccenting and durational cues (Guaraní). The languages with weaker macro-rhythm used accenting and phrasing as focus marking cues, which result in more variable f0 patterns (English).

Other research investigated the relationship between macro-rhythm and non-linguistic aspects. One study used Mac*R*_Var to quantify speech produced in different acted emotional states (Arronte Alvarez et al. 2022). The results showed that high-arousal emotional states (e.g., anger, surprise) lead to more macro-rhythmic speech than low arousal emotional states (e.g., sadness). Another study, the perception of the degree of Jewishness in Jewish English stimuli by Jewish and non-Jewish listeners was investigated with a particular focus on the shape of fo contours (Burdin 2020). It was found that higher degrees of macro-rhythm (more regular alternations of H and L tones) was the cue that both listener groups associated with higher degrees of Jewishness.

In sum, the results of Polyanskaya et al. (2020) and Prechtel (2020b) showed that, interestingly, purely tonal metrics captured the expected macro-rhythmic differences best (Table 2). This is an important observation in rhythm research, which has been devoted to the investigation of temporal regularity for decades (e.g., Fletcher 2010). It is particularly noteworthy, because Mac*R*_Var was proposed specifically as *the* quantification of macro-rhythm and failed to show typological differences in both studies, and only successfully distinguished emotional states (Arronte Alvarez et al. 2022). As noted in Polyanskaya et al. (2020), more research is needed on the tonal component as independent contributor to (macro-)rhythm

as a typological notion. This issue is taken up in the current study and further outlined in the next section.

1.4 A non-timing tonal approach to macro-rhythm

Before turning to the research aims of the current study, some important remarks should be made given the literature discussed in the previous sections. In particular, it can be questioned what it means for languages to be more or less macro-rhythmic in light of the language differences found in rhythm perception (Barry et al. 2009; Cumming 2011). There are two possible implications for macro-rhythm. First, it might be more realistic to interpret the degree of macro-rhythmicity as the degree to which f0 plays a role in rhythm perception among the other possible acoustic cues in a given language (e.g., the degree of *tonal rhythmicity*). Second, the notion macro-rhythm could be expanded such that it includes all possible acoustic correlates to rhythm (i.e., termed *micro* rhythm in Jun 2005). Note, however, that for the majority of the languages in Jun (2014) the relative importance of the acoustic cues to rhythm is unknown. The typology is largely based on expectations based on existing research on prosodic structure, in particular research on word- and phrase-level prominence marking and autosegmental-metrical accounts based on inventories of pitch accents and/or phrasal tones of AP-like prosodic units. Furthermore, the dependency between duration and f0 in rhythm differs per language (Cumming 2011).

As called upon in the literature, it is time to approach rhythm without prioritizing the search for regularity and explore new ways of rhythm quantification. The notion of macro-rhythm in particular demands more research on the tonal component in rhythm. A logical next step, therefore, is to quantify macro-rhythm using a metric that does not rely on timing measures. Thus, the aim of the present study is to find out how successful macro-rhythm quantification, based on f0 shape only, is in typologizing languages. This not only provides a new perspective to tonal rhythm, it also has the potential to offer a novel way to investigate the prosodic structure of under-researched languages. The latter merit is particularly welcome for language documentation purposes, for two reasons. First, macro-rhythmic research so far is largely based on well-researched languages (Table 2). Second, the metric proposed in this chapter can be a first step in documenting prosodic characteristics of a language for which there are little to no scientific (re-)sources. The research aims are further specified in the following section.

1.5 Research aims

The discussion above indicates that it could indeed be a promising approach to quantify macro-rhythm using f0 shape metrics only. Section 1.2 argues that macro-rhythm depends mainly on the surface realization, in particular on the type and frequency of tonal targets, rather than on underlying accent inventory size. Note that this is a simplified and non-specific view for the purposes of the current study and is not meant to deny any contribution of temporal regularity to the notion of rhythm. The simplification in the current study is justified as it offers a novel theoretical approach to rhythm and a feasible metric for language documentation purposes. Thus, to capture both the frequency and the width of f0 excursions, the current study proposes a single metric of f0 contour length. This metric treats the f0 contour as if it were a string that is bent. That is, an f0 contour with many and wide excursions makes up a longer "string" than a contour with less and shallow excursions. Two theoretical extremes are schematized in Figure 1 and show that a contour length could be approximately one and a half times as long with frequent and wide excursions than without any excursions. The extremes in Figure 1 are unlikely to be produced in natural language, so more subtle differences in contour length are expected among languages that differ in macro-rhythmicity. Note that this way of measuring f0 contours — to my knowledge — has not been reported in the literature before and offers a new way of investigating intonation contours. Contour length, of which the precise measuring is explained in Section 2, provides a coarse indication of the amount of "tonal activity" without reliance on the exact meaning of each f0 movement. That is, it is insensitive to prosodic structure in terms of pre-annotated tonal targets at the word- or phrase-level (as used in Polyanskaya et al. 2020 and Prechtel 2020b) and does not need assumptions about accent inventory to be operationalized (cf. Section 1.2). The latter benefit enables researchers to document coarse prosodic characteristics of a language without having to resort to resource-intensive tonal annotations. It therefore offers a way to test the following hypothesis:

> The degree of macro-rhythm can be approximated using a metric based on contour length, with longer contours corresponding to higher degrees of macro-rhythm.

The hypothesis is tested using a contour length metric applied to speech from five languages; Greek, German, European Portuguese, French and Seoul Korean. These languages were chosen because they are claimed to have different degrees of macro-rhythm (Table 1). Particular variation in macro-rhythm is expected among the head languages; Greek (strong), German (medium) and European Portuguese (weak). French (head/edge language) and Seoul Korean (edge language) are

listed as strongly macro-rhythmic (such as Greek), although the degrees among the strongly macro-rhythmic languages might still differ (Jun 2014: 533). Table 3 lists all the relevant prosodic characteristics for the degree of macro-rhythm as reported in Jun (2014).

The languages investigated in this study are chosen as the literature offers existing and testable claims on their degree of macro-rhythmicity (Table 1). It should be noted, however, that the simplified and generalized contour length metric lends itself particularly well for application in the documentation of under-researched languages. This holds in particular for languages for which little to no annotated speech is available. The minimal requirement for the metric to be applied is segmentation into (intonational) phrases and availability of fo measurements. The discussion (Section 4) will return to the usefulness of this metric for under-researched languages.

Table 3. Prominence type, word prosody, macro-rhythm degree, accent inventory size (N_{acc}), most common accent (Mo_{acc}) and domain of head/edge per language (incl. ISO codes); following Jun (2014)

Language	ISO	P.type	W. prosody	Macr.R	N_{acc}	Mo_{acc}	Domain
Greek	ELL	head	stress	strong	5	L*+H	≈ Cwd
German	DEU	head	stress	medium	6	H*	≥ Cwd
European Portuguese	POR	head	stress	weak	5	H*/(H)	> Cwd
French	FRA	head/edge	none	strong	4	LH*	≥ Cwd
Seoul Korean	KOR	edge	none	strong	6	LH	≥ PW

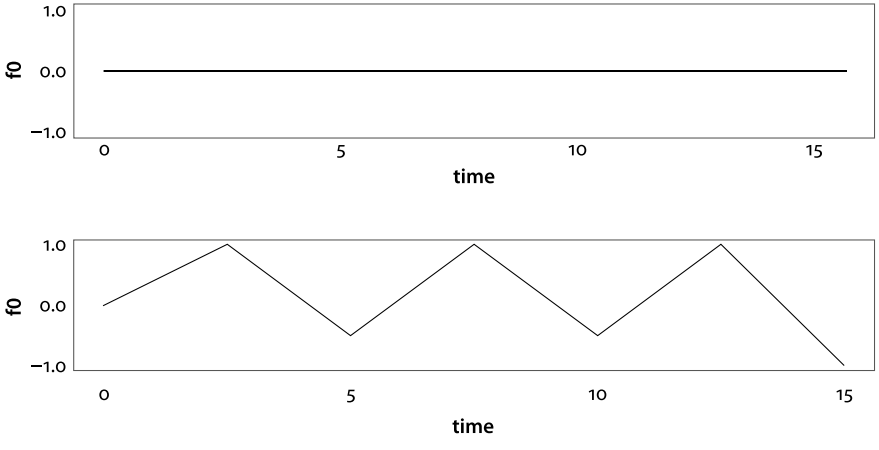


Figure 1. Two theoretical extremes of the degree of macro-rhythm; weak (top) and strong (bottom) with arbitrary values for time (x-axis) and fo (y-axis)

2. Methodology

Two acoustic analyses were carried out. In the first exploratory analysis short readings of the Aesop fable "The North Wind and the Sun" (e.g., Joseph 1894) were analyzed (scripted speech from one speaker per language), whereas in the second more confirmatory analysis carefully selected phrases from corpora of TED Talks (TED Conferences LLC 1984) were analyzed (from multiple speakers per language). The scripted speech allows for control over the contents of the utterances, i.e., the story is identical in each language. However, the stories are short and are not a representative sample of natural language. The latter are taken as a more spontaneous type of speech data. That is, although the TED talks are generally rehearsed, they are delivered in an unscripted way. The analysis of unscripted speech was carried out using Bayesian hierarchical linear models fitted on the acoustic measurements. This section reports further methodological details for each of the analyses.

2.1 Scripted speech

Recordings of the North Wind and the Sun were obtained from an online speech archive (De Iacovo 2017) for each of the languages. They consist of laboratory recordings of a native speaker (young to middle-aged) reading the fable. All recordings were classified as high quality (De Iacovo 2017). Additional information per recording is provided in Table 4 and transcripts are provided in Appendix I.

Each phrase in the recording was annotated in Praat (Boersma & Weenink 2022). Phrase boundaries corresponded to dots in the transcripts and delimited either five or six phrases depending on the language (Table 4). Silent intervals between phrases were not included. The phrase boundaries were saved in TextGrids (Boersma & Weenink 2022).

Table 4. Recording information per language/speaker (scripted speech: The North Wind and the Sun)

Language	ISO	Gender	Origin	N phrases
Greek	ELL	M	Athens	6
German	DEU	M	Bielefeld	6
European Portuguese	POR	F	Lisbon	5
French	FRA	F	Rouen	6
Seoul Korean	KOR	F	Seoul	5

2.2 Unscripted speech

For the analysis of the unscripted speech, two corpora of TED talks were used. The former consisted of talks in Greek, German, European Portuguese and French (Salesky et al. 2021), whereas the latter consisted of Seoul Korean ones. The Seoul Korean talks were obtained from Youtube recordings and their accompanying transcript files (see Appendix II for an overview of all the videos). The accompanying transcript files were converted into TextGrids. The transcripts were based on TED community (TED Conferences LLC 1984) translations and automatically aligned subtitles based on automatic speech recognition models (e.g., Gulati et al. 2020) as implemented in Youtube (Google LLC 2022). Based on the transcripts, a number of criteria was applied to select phrases that were suitable for a cross-language comparison. That is, only phrases ending with a dot (.) in the transcription were selected. Question marks (?) or commas (,) were excluded to avoid an effect of non-finality on the final part of the f0 contour. Note that questions and (intermediate) phrases often have their own, potentially rising, final intonation, which would jeopardize the comparability to phrase endings with a low or falling fo. The dots ensured a final low/fall in all languages, which was confirmed by a manual check after selection. Furthermore, phrases including commas (,), colons (:) or semi-colons (;) in non-final position were also excluded from analysis. This was done to avoid larger pauses (and potentially tonal resets) within a phrase, which would affect the fo contour to a larger extent than phrases without a (comma-)break. In addition, phrases that did not start with a capital letter (Korean transcripts were translated into English for this purpose) were excluded to further avoid analysis of incomplete phrases.

The remaining set of phrases was furthermore reduced to the ones that fit a duration range of 200 ms. around the median duration of that subset (2.75 seconds). This range ensured that the selected intervals were of similar length such that the amount of possible speech and intonation events in the interval were comparable across the languages. A comparison of number of words would not have accounted for morphological differences between the languages. An absolute duration criterion was therefore estimated as the most equal way of comparing the languages. The mean number of words per phrase were calculated for each language and given for reference in Table 5. Note that the Korean corpus was much smaller than the ones for the other languages. This reduced the number of speakers and phrases that could be analyzed for this language. Table 5 reports the number of speakers, phrases and mean number of words per phrase that were ultimately analyzed statistically (after further processing, see Section 2.3 and Section 2.5). Given that the Korean TED Talks were not obtained from the corpus that was used for the other languages, all selected Korean phrases were addi-

tionally checked and it was ensured that they had good audio quality and that the speech was indeed coming from the speaker (not from any audio played in/added to the talk).

Table 5. Recording information per language (unscripted speech: TED talks)

Language	ISO	N speakers	N phrases	M words/phrase (SD)
Greek	ELL	41	87	6.33 (2.27)
German	DEU	55	146	6.01 (2.05)
European Portuguese	POR	46	91	6.54 (2.15)
French	FRA	35	62	7.10 (2.43)
Seoul Korean	KOR	16	41	6.76 (2.68)

2.3 Acoustic measurements

Each recording and accompanying TextGrid were used to obtain time-series f0 measures (both the scripted and unscripted data). That is, for each phrase interval 40 f0 measurements were taken in equally distributed intervals over the phrase. This was done by dividing the duration of the phrase into 40+1 equidistant intervals, such that measurement point 1 and measurement point 40 were both equally far away from the phrase edges as from a neighboring measurement point. Note that f0 valleys or peaks are not precisely measured at their turning points using this method. However, given the mean phrase length (Section 2.2), this method ensured that one measurement point was taken at approximately each 69 ms. window (equals two to five measurement points per syllable), which is estimated as a sufficient amount to capture all relevant f0 movements in the contour. The "Advanced pitch settings" in Praat (Boersma & Weenink 2022) were set to values that captured the f0 track best in these recordings and are reported in Table 6. Most of the settings were left at their standard values, except Voicing threshold and Octave-jump cost. Both values were set to higher ones to remove inaccurately tracked f0 at the edges of contours (Voicing threshold) and to avoid octave jumps that could result in inaccurate f0 tracking as well (Octave-jump cost). The obtained f0 measures also included duration information; the start and end time of the phrase interval and the duration between the measurement points/phrase edges (a constant for each phrase). These duration measures were used to calculate contour length, as explained in the following subsection.

The success of the octave-jump correction was assessed by calculating the mean f0 for each phrase before and after octave-jump correction. If these values

Table 6. Advanced pitch settings used for the time-series f0 measures

Setting	Value used	Standard value
Minimum pitch	75	(idem)
Maximum pitch	500	(idem)
Silence threshold	0.03	(idem)
Voicing threshold	0.55	0.45
Octave cost	0.01	(idem)
Octave-jump cost	0.55	0.35
Voiced/unvoiced cost	0.14	(idem)

differed more than 10%, the contour was discarded. This procedure reduces the number falsely corrected octave-jumps and is based on Kaland (2021).

The f0 measures were furthermore speaker normalized using octave-median (OMe) rescaling (De Looze & Hirst 2014, see Equation 5). This rescaling uses the octave as a unit for speech melody and was shown to be an accurate estimation of the natural fo range of a speaker. A preliminary version of this study showed that octave-median rescaling outperformed a more generic speaker-normalization method based on robust rescaling for the use of contour length measures (Kaland 2022).

The unscripted data was furthermore subset to reduce measurement imprecision as a result of the alignment between transcript and speech. The automatically generated transcripts potentially included silent intervals before or after the phrase. Contours for which the first or last 20% had no movement were therefore discarded. This was calculated by taking the standard deviation (σ) of the initial eight and final eight measurement points. If the σ was found zero at either of these eight points, the contour was not taken into account for further analysis. For the unscripted speech data additional calculations of f0 range per language were taken by subtracting the highest rescaled f0 from the lowest rescaled f0 among all measures for that language. This measure was omitted for the scripted speech due to the small amount of data.

$$\text{OMe} = \log 2\left(\frac{Hz}{median}\right) \quad (5)$$

2.4 Contour length calculation

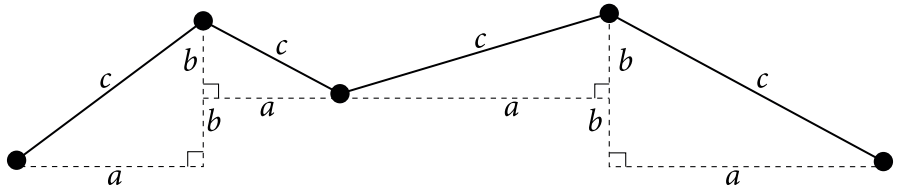


Figure 2. Example of a stylized f0 contour (solid line) with 5 measurement points (dots) and with contour length (sum of all *c*'s) as calculated by Pythagoras' theorem, illustrated using right triangles (dotted lines)

Contour length was approximated using Pythagoras' theorem and is illustrated in Figure 2. In this calculation *a* is the duration difference and *b* is the f0 difference between two consecutive measurement points. Given *a* and *b, c* can be calculated and approximates the length of the f0 contour between two measurement points. For the length of the entire contour (L_c), the sum of all *c*'s was taken. As such, L_c is not a meaningful measure given that the duration of phrases varied; either because they were uncontrolled for (Section 2.1) or because there was a small degree of variation within a range of 400 ms. centered around the median (Section 2.2). Therefore, contour length is more informative when knowing the exact length of the phrase. For this reason, the increase in length was calculated relative to a flat line spanning all measurement point intervals (L_i), which equals the sum of all *a*'s. Recall the two theoretical extremes in Figure 1. Contour length increase (CLI) therefore expresses the relative length difference between interval length (L_i) and the f0 contour of that interval (L_c) and captures how much longer a the f0 contour is compared to when that contour would have been a (monotonous) flat line. The resulting CLI values, in particular when based on rescaled f0 values are extremely small and therefore expressed as a parts per million (multiplied by 1,000,000) for the ease of interpretation and reporting (following a comparable method in Kaland 2022). The equation is given in Equation 6.

(6) $$\text{contour length increase (CLI)} = 1\,000\,000 * \left[\frac{L_c - L_i}{L_i}\right]$$

2.5 Statistical analysis

Statistical analysis was performed on the unscripted speech only, as this dataset was large enough to yield representative results. Before analysis, outlier removal was performed on the CLI values per language. This was done to avoid effects of extreme values on the statistical analysis. Manual inspection of the outliers confirmed that they were the result of remaining f0 tracking errors, which resulted in extreme contour length values. The number of outliers per language was as follows: Greek: 1, German: 4, European Portuguese: 6, French: 5, Seoul Korean: 2. Outliers were defined as the CLI values that lied outside 1.5 times the inter-quartile range (IQR) above the upper quartile (Q3) as approximated by: [Q3 + 1.5 * IQR].

One-sided hypotheses on the degree of macro-rhythmicity following the ones in Jun (2014) were formulated for the head languages, comparing all of them to each other: ELL > DEU, ELL > POR, DEU > POR. Furthermore, two-sided hypotheses were formulated for the supposedly strongly macro-rhythmic languages (expected to have comparably strong macro-rhythm in Jun, 2014): ELL = FRA, ELL = KOR, FRA = KOR.

Bayesian hierarchical linear models ('brms' package; Bürkner 2017, using 'Stan'; Carpenter et al. 2017) were fitted to the CLI values (gamma distribution) in R (R Core Team 2020) and R Studio (R Studio Team 2020) with language (5 levels: ELL, DEU, POR, FRA, KOR, reference level ELL) as predictor and speaker as random intercept, using a defined prior for all population-level effects (class 'b'): a normal distribution with $\mu = 0$ and $\sigma = 1$. These values were chosen because the model fitted best with a log-linked gamma distribution, i.e., a mean difference set to 0 is a conservative estimate of no differences between the languages and a standard deviation of 2 is a non-specific estimate covering 98.59% of the data. Default priors were used for standard deviations of random effects (a Student's t-distribution with $\nu = 3$, $\mu = 0$ and $\sigma = 2.5$). Due to the inclusion of two-sided hypotheses, priors were also sampled and in order to avoid divergent transitions the number of iterations was set to 4000 with 2000 warm-up iterations, drawn using the Hamiltonian Monte Carlo algorithm (e.g., Neal 2012) in four chains (total post-warmup draws = 8000).

The expected values under the posterior distribution and their 95% credible intervals (CIs) are reported for each hypothesis (Table 9). The posterior probability that a hypothesized difference is bigger than zero is also reported. The results are interpreted as confirming the hypothesis if zero is not included in the 95% CI of the hypothesized difference and the posterior probability is close to one.

3. Results

3.1 Scripted speech

The results of the scripted speech show that Greek has the highest CLI values, whereas European Portuguese shows the smallest values (Table 7). These values need to be interpreted as a rough indication of contour length increase, i.e., based on a small set of phrases. Examples of phrases in each language are given in Figure 3. The overall ordering according to CLI shows some further differences between the languages. Among the head languages, the CLI in Greek is higher than in German, which in turn is higher than in European Portuguese. Among the supposedly strong macro-rhythmic languages with difference prominence types, there is however considerable difference between Greek (strong), French (medium) and Seoul Korean (weak).

It can furthermore be observed from the mean phrase lengths (Table 7), that French and Korean differed considerably from the other languages in this respect. These considerable differences are likely to have some effect on the f0 measures, in particular given that 40 measurement points give a different measurement resolution to a phrase with a length of 2.12 seconds (the minimum in the data) compared to a phrase with a length of 10.24 seconds (the maximum in the data). In other words, some f0 movements might have been missed in longer phrases because the intervals between the measurement points were longer. Further discussion of these results is given in Section 4.

Table 7. Mean contour length increase (CLI), mean phrase length in seconds, and standard deviations per language in the scripted speech

Language	ISO	CLI	Phrase length (s)
Greek	ELL	0.69 (0.34)	6.27 (2.29)
German	DEU	0.62 (0.20)	6.22 (2.56)
European Portuguese	POR	0.25 (0.06)	6.31 (1.95)
French	FRA	0.51 (0.43)	5.08 (2.39)
Seoul Korean	KOR	0.32 (0.05)	8.20 (1.32)

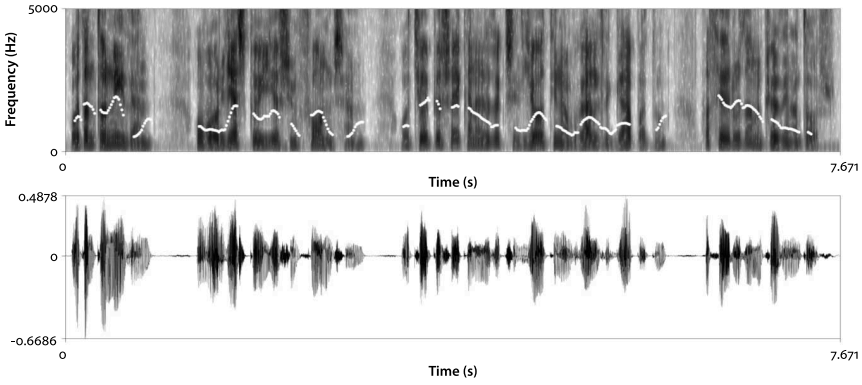


Figure 3a. Greek: *Όταν τον είδαν, ο βοριάς κι ο ήλιος συμφώνησαν ότι όποιος έκανε τον ταξιδιώτη να βγάλει την κάπα του θα θεωρούνταν ο πιο δυνατός*

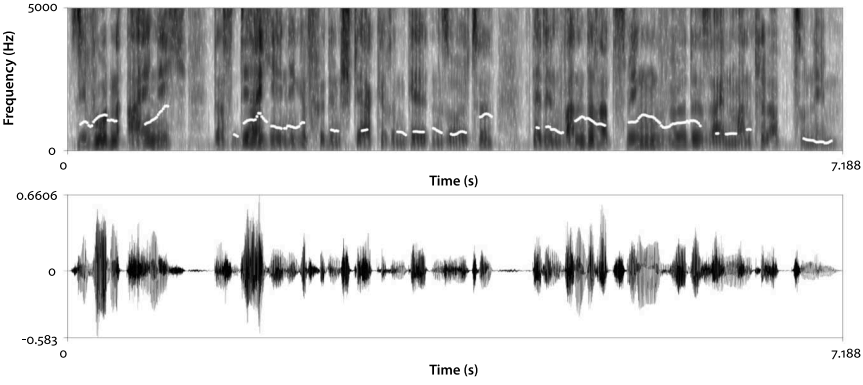


Figure 3b. German: *Sie wurden einig, daß derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen*

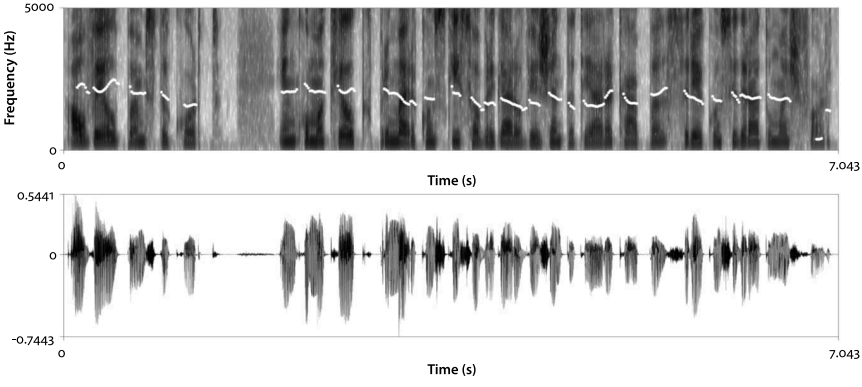


Figure 3c. Portuguese: *Ao vê-lo, põem-se de acordo em como aquele que primeiro conseguisse obrigar o viajante a tirar a capa seria considerado o mais forte*

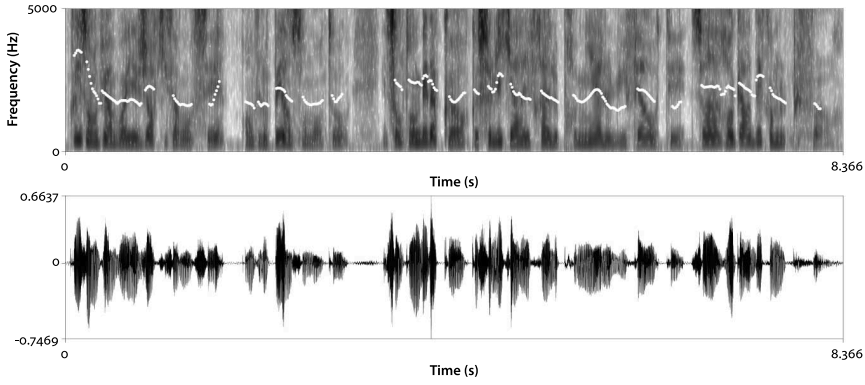


Figure 3d. French: *Quand ils ont vu un voyageur qui s'avançait, enveloppé dans son manteau, ils sont tombés d'accord que celui qui arriverait le premier à le lui faire ôter serait regardé comme le plus fort*

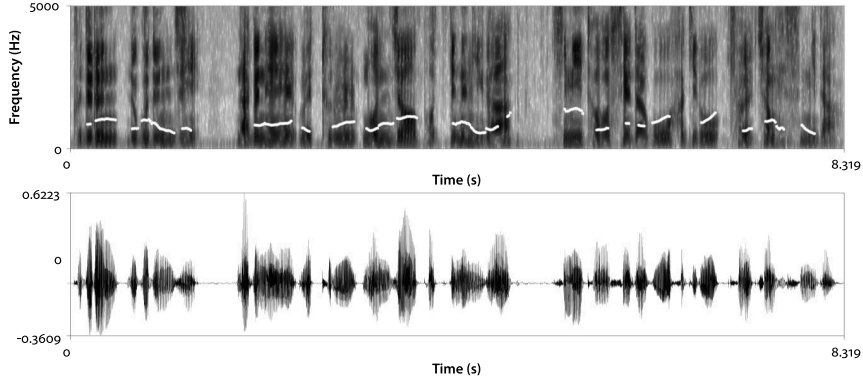


Figure 3e. Korean: 그들은 누구든지 나그네의 외투를 먼저 벗기는 이가 힘이 더 세다고 하기로 결정했습니다

Figure 3. Examples of the second phrase in the North Wind and the Sun: *They agreed that the one who first succeeded in making the traveler take his cloak off, should be considered stronger than the other*

3.2 Unscripted speech

The results of the statistical analysis on the TED talks in each language are given in Table 8, Table 9 and Figure 4. Among the head languages Greek had the largest mean CLI, which differed significantly from the one of German and European Portuguese. The mean CLI of German did not differ significantly from the one of European Portuguese, with the posterior probability of 0.55 indicating

a near chance-level probability that the two languages had a different contour length deviation. Among the strongly macro-rhythmic languages, Greek also had the largest mean CLI, followed by French and Seoul Korean. No significant differences could be confirmed among the strongly macro-rhythmic languages, although the posterior probabilities show that Greek and French as well as French and Korean were highly likely to be similar.

Figure 4 indicates notable differences between the distributions of the CLI values among the languages. That is, Greek shows the largest range of CLI values, with most of the values found around 0.6 and a minimum of 0.10. German (minimum 0) and European Portuguese (minimum 0.02) show a more compressed violin shape with smaller CLI ranges and lower modes (~0.40). The difference between German and European Portuguese is most obvious from the modes; European Portuguese has a more uni-modal distribution than German. French shows a uni-modal distribution of the CLI values with the minimum value at 0.18. Seoul Korean CLI values have a mode around 0.40 and a minimum of 0.04. The CLI range of Seoul Korean is similar to the one of European Portuguese although the mode is lower in the former (around 0.25).

Table 8. Expected degree of macro-rhythm (following Jun, 2014), mean contour length increase (CLI) and standard deviation, minimum CLI, mean phrase length in seconds and standard deviation, and fo range (OMe rescaled) per language in the scripted speech

Language	ISO	Macr.R	*M* CLI	Min. CLI	Phr. length (s)	F0 range
Greek	ELL	strong	0.87 (0.59)	0.10	2.74 (0.10)	3.79
German	DEU	medium	0.71 (0.47)	0.00	2.76 (0.12)	3.13
European Portuguese	POR	weak	0.69 (0.40)	0.02	2.74 (0.12)	2.35
French	FRA	strong	0.82 (0.48)	0.18	2.74 (0.12)	3.30
Seoul Korean	KOR	strong	0.66 (0.42)	0.04	2.77 (0.10)	2.39

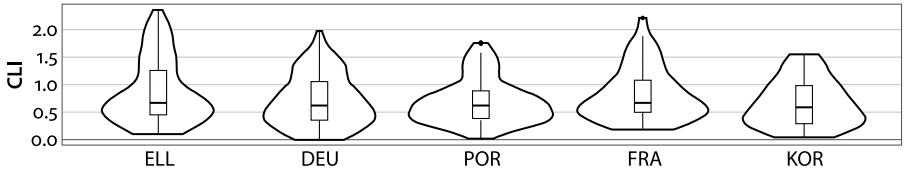


Figure 4. Violin plots of contour length increase (CLI) per language

Table 9. Hypothesis tests on the Bayesian model

Lg. group	CLI hyp.	Estimate	Est. error	CI. lower	CI. upper	Post. prob.	Sign.
Head lgs	ELL>DEU	0.22	0.11	0.05	0.39	0.98	*
	ELL>POR	0.23	0.11	0.04	0.41	0.98	*
	DEU>POR	0.01	0.11	−0.16	0.18	0.55	
Strong macr.r.	ELL=FRA	0.06	0.12	−0.18	0.31	0.87	
	ELL=KOR	0.25	0.15	−0.05	0.53	0.62	
	FRA=KOR	0.18	0.16	−0.13	0.49	0.81	

4. Discussion and conclusion

The results of this study have shown that contour length deviation is an informative measure to quantify the degree of macro-rhythm. This outcome particularly holds for head languages, which were expected to show most variation in the degree of macro-rhythm. Although both the scripted and the unscripted data confirmed the ordering of head languages from strong to weak (ELL > DEU > POR), the Bayesian analysis of the unscripted data did not show significant differences between German and European Portuguese. As for the variation within the strong macro-rhythm group (ELL, FRA, KOR), it was shown that Greek was similar to French and French was similar to Korean. Thus, the hypotheses could be confirmed to some extent for the strongly macro-rhythmic languages. Nevertheless, there is still some variation among the mean CLI values of these languages. Some differences among the strongly macro-rhythmic languages could be expected (Jun 2014: 533), although the CLI for Seoul Korean (strong) turned out to be lower than the ones for European Portuguese (weak). The latter finding suggests that there is a limitation to the mapping of the CLI metric to macro-rhythmic differences. In this respect it is important to note that for each prominence type group of languages, the degree of macro-rhythm is expected to vary to a different degree (e.g., Var(head lgs) > Var(head/edge lgs) > Var(edge lgs); see Table 1). Thus, a CLI of 0.66 (KOR) might indeed be indicative of strong macro-rhythm among the edge languages, although it does not reconcile with a CLI of 0.87 as found in a strong macro-rhythmic head language (ELL). Both the scripted and the unscripted data also revealed the same ordering of languages with strong macro-rhythm (ELL > FRA > KOR), suggesting that this result is due to linguistic differences in prosodic structure (i.e., prominence type), see Table 7 and Table 8. More languages of each prominence type need to be investigated to confirm this.

There appears to be a shared feature among the strongly macro-rhythmic languages tested in this study in that they have higher minima in the CLI distribu-

tions (ELL, FRA, KOR), see Figure 4. This observation not only contributes to the explanation of the higher mean CLIs in the Greek and French unscripted data, it also indicates that the languages differed in the amount of minimal tonal movement in a phrase. Inspection of Table 3 indicates that this could be the result of the shape of the most common accent. That is, Greek, French and Seoul Korean all have bi-tonal accents among the most common ones (L and H combined in that order), whereas German and European Portuguese mostly apply singleton accents (H). Thus, applying accents with multiple tonal targets will result in contour lengths that deviate more than applying accents with one tonal target. This result means that small length deviations (e.g., < 0.18 for French) did not occur in the unscripted data of strongly macro-rhythmic languages. Note that this phenomenon is essentially different from the overall f0 range differences between languages, although some correlation is expected (Table 8).

There are two further questions to the outcomes of this study. First, what does it mean for the notion of macro-rhythm if it can be captured relatively well by a metric excluding a timing component? And second, which other tonal aspects are captured by CLI that might be relevant for the (typological) study of prosody? Both questions relate to the nature of the CLI metric and are discussed in the following.

Although the CLI metric was informative to some extent, its application in this study is rather exploratory. It was speculated that overall tonal activity correlates with the "opportunity" of perceiving rhythmicity from f0 variations (Section 1.5). The CLI values were based on production data and could therefore not confirm this directly, but still suggest that it is a plausible approach. Note that CLI was proposed as a simplified approach to macro-rhythm quantification. The literature has shown that both tonal and temporal factors contribute to perceived rhythm (Barry et al. 2009; Cumming 2011), and the results of this study should be interpreted in the light of these findings. That is, they are in need of further refinement using temporal (and potentially other) acoustic metrics. In this respect, it is also important to note that the f0 measures concern converted values using the OMe rescaling (De Looze & Hirst 2014). This rescaling took into account the individual speaker's f0 range and accounted for melodic differences in an accurate way. The rescaling resulted in extremely small contour length increase values (see Section 2.4). Note that the examples given in Figure 1 and Figure 2 where an increase factor of approximately 1.5 (e.g., 150%) could be observed purely served illustration purposes and are not comparable to the OMe rescaled CLI values. The latter were expressed as parts per million, which was considered a more realistic approach to the length differences. Given the almost identical phrase lengths in the unscripted data across the five languages (Table 8), the way CLI was measured was indeed able to reveal relevant language differences, to an extent that matched the expected macro-rhythm differences to a considerable extent. It is also impor-

tant to note in this context that the duration scale was left unconverted, which is expected to affect the CLI calculation as well. The calculation of CLI could furthermore be refined by comparing contour length to a declining contour instead of a flat contour, as f0 movements are produced and perceived relative to the baseline fo following the declination effect (Pierrehumbert 1979; 't Hart et al. 1990, Chapter 5). These refinements as well as the combination of contour length with temporal (timing) metrics are left for future work.

Perception studies are another way in which future work could contribute to the current results. If it holds that listeners can perceive isochrony in a signal from which it is strictly speaking not measurable (as suggested in the literature discussed in Section 1.1), the question arises to what extend speech rhythm needs a highly regular component in the signal to begin with. While it is well known that phrase- and word-boundaries are signaled prosodically in many languages (e.g., Cutler 2012), these boundaries are not necessarily found in regular intervals. It seems therefore, that an important task in speech perception (chunking the incoming signal) can be done when its cues appear irregularly. This observation does not deny that any form of *perceived* regularity is not helpful to listeners. It would mean that rhythm is a merely perceptual construct (as suggested in Arvaniti 2009) that depends on the amount of "opportunity" the signal provides to hear it (see this argument outlined in Section 1.5). This view is compatible with the way contour length has been analyzed in this paper, i.e., as a measure of overall tonal activity increasing the likelihood that a rhythm can be perceived.

How exactly temporal aspects of the signal play a role in perceived macro-rhythm still remains to be investigated. The current results are promising in that they shed a new light on the potential of a purely tonal generic metric to rhythm. They also provide a partial answer to the questions that were raised on the definition of macro-rhythm (Section 1.4). That is, macro-rhythm mostly captures the tonal aspect of rhythm, and should thus be seen as an incomplete speech rhythm metric. In similar vein, this study has provided a metric that should also be further refined using complementing metrics based on temporal aspects of speech. For example, CLI would not be able to distinguish phrases in which f0 movements are crowded at one location from phrases with more distributed f0 movements. The latter type would be predicted to be more macro-rhythmic than the former in Jun (2014). A very specific hypothesis brought forward by this study pertains to the varying CLI measures among strongly macro-rhythmic languages. It can be expected that languages of different prominence types differ in the extent to which they rely on temporal speech characteristics for the perception of rhythm. This might hold in particular for the languages with edge tones (head/edge and edge languages), because these tones tend to take a larger domain for realization (shallower slopes) compared to pitch accents (found in head languages). It

thus remains to be seen how the cross-linguistic differences turn out after adding temporal metrics to the CLI. More in general, still more languages, in particular non-Indo-European ones, should be included in future work to complement the results obtained in this study. The applicability of CLI to under-documented languages is of particular interest in this context. As briefly mentioned in Section 1.5, CLI requires mainly phrase-level boundaries and f0 measurements in order to be computed. Elaborate annotations, often too resource-intensive at early stages of language documentation, are not required. While these aspects mainly offer practical advantages for the application of CLI, it is still challenging to use this metric reliably to prosodically typologize under-researched languages. This is the case due to at least two important issues. First, without further knowledge of the prosodic characteristics, applying a purely tonal metric cannot be conclusive for any kind of prosodic typological classification, whether macro-rhythmic or any other dimension. Second, the type of speech data collected in fieldwork settings differs from the TED talks tested here. While these talks might be representative to some degree of the languages investigated in this study, they will not be available for languages spoken in geographically remote areas by small communities. The question remains whether other types of speech data are suitable to apply CLI. The exploratory analysis of the North Wind and Sun readings seems to suggest so, even though these kinds of readings are not representative of languages for which little to no literary tradition exists. Regardless of these future challenges, the current results may be interpreted as a low-threshold invitation to apply the CLI metric to a wider variety of languages and speech types, in particular under-documented ones.

Funding

Research for this paper was funded by the German Research Foundation (DFG) — Project-ID 281511265—SFB 1252.

Acknowledgements

A previous version of this study was presented at Interspeech 2022 (Kaland 2022). The author thanks Sun-Ah Jun for valuable comments on this study. Supplementary materials for this study can be found at: https://osf.io/rd8gx/.

References

Arronte Alvarez, Aitor, Elsayed Issa, and Mohammed Alshakhori. 2022. "Computational Modeling of Intonation Patterns in Arabic Emotional Speech." In *Proceedings of Speech Prosody 2022*, 615–619. Lisbon, Portugal.

Arvaniti, Amalia. 2009. "Rhythm, Timing and the Timing of Rhythm." *Phonetica* 66 (1–2): 46–63.

Arvaniti, Amalia. 2012. "The Usefulness of Metrics in the Quantification of Speech Rhythm." *Journal of Phonetics* 40 (3): 351–373.

Arvaniti, Amalia, and Mary Baltazani. 2005. "Intonational Analysis and Prosodic Annotation of Greek Spoken Corpora." In *Prosodic Typology*, edited by Sun-Ah Jun, 84–117. Oxford: Oxford University Press.

Barry, William, Bistra Andreeva, and Jacques Koreman. 2009. "Do Rhythm Measures Reflect Perceived Rhythm?" *Phonetica* 66 (1–2): 78–94.

Boersma, Paul, and David Weenink. 2022. "Praat: Doing Phonetics by Computer." http://www.praat.org/

Burdin, Rachel Steindel. 2020. "The Perception of Macro-Rhythm in Jewish English Intonation." *American Speech* 95 (3): 263–296.

Burdin, Rachel Steindel, Sara Phillips-Bourass, Rory Turnbull, Murat Yasavul, Cynthia G. Clopper, and Judith Tonhauser. 2015. "Variation in the Prosody of Focus in Head- and Head/Edge-Prominence Languages." *Lingua* 165 (October): 254–276.

Bürkner, Paul-Christian. 2017. "Brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1).

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

Cumming, Ruth E. 2011. "The Language- Specific Interdependence of Tonal and Durational Cues in Perceived Rhythmicality." *Phonetica* 68 (1–2): 1–25.

Cutler, Anne. 2012. *Native Listening: Language Experience and the Recognition of Spoken Words*. Cambridge, Mass: The MIT Press.

Dauer, R. M. 1983. "Stress-Timing and Syllable-Timing Reanalyzed." *Journal of Phonetics* 11 (1): 51–62.

De Iacovo, Valentina. 2017. "Speech Archive — The Northwind and the Sun." LFSAG — Laboratorio Di Fonetica Sperimentale "Arturo Genre." 2017. https://www.lfsag.unito.it/ark/table_ldm_en.html

De Looze, Céline, and Daniel Hirst. 2014. "The OMe (Octave-Median) Scale: A Natural Scale for Speech Melody." In *Proceedings of Speech Prosody 2014*, 910–914. ISCA.

Dellwo, Volker, Adrian Fourcin, and E. Abberton. 2007. "Rhythmical Classification of Languages Based on Voice Parameters." In *Proceedings of the 16th International Congress of Phonetic Sciences*, edited by Jürgen Trouvain, and William Barry, 1129–1132. Saarbrücken, Germany.

Fletcher, Janet. 2010. "The Prosody of Speech: Timing and Rhythm." In *The Handbook of Phonetic Sciences*, edited by William J. Hardcastle, John Laver, and Fiona E. Gibbon, 521–602. Oxford: Blackwell.

Google LLC. 2022. "YouTube." 2022. https://www.youtube.com/

Gordon, Matthew K. 2014. "Disentangling Stress and Pitch-Accent: A Typology of Prominence at Different Prosodic Levels." In *Word Stress*, edited by Harry van der Hulst, 83–118. Cambridge: Cambridge University Press.

Gordon, Matthew K., and Timo Roettger. 2017. "Acoustic Correlates of Word Stress: A Cross-Linguistic Survey." *Linguistics Vanguard* 3 (1): 20170007.

Grabe, Esther, and Ee Ling Low. 2002. "Durational Variability in Speech and the Rhythm Class Hypothesis." In *Laboratory Phonology 7*, edited by Carlos Gussenhoven and Natasha Warner. Berlin and Boston: De Gruyter Mouton.

Grice, Martine, Stefan Baumann, and Ralf Benzmüller. 2005. "German Intonation in Autosegmental-Metrical Phonology." In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by Sun-Ah Jun, 55–83. Oxford: Oxford University Press.

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zang, Yonghui Wu, and Ruoming Pang. 2020. "Conformer: Convolution-Augmented Transformer for Speech Recognition."

Hart, Johan 't, René Collier, and A. Cohen. 1990. *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Hirst, Daniel J. 2007. "A Praat Plugin for Momel and INTSINT with Improved Algorithms for Modelling and Coding Intonation." In *Proceedings of the 16th International Congress of Phonetic Sciences*, 1233–1236. Saarbrücken, Germany.

Joseph, Jacobs. 1894. *The Fables of Aesop*. Vol. 60, pp. 142–143. London: Macmillan & Company.

Jun, Sun-Ah (ed.). 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford Linguistics. Oxford: Oxford University Press.

Jun, Sun-Ah (ed.). 2014. *Prosodic Typology II: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, Sun-Ah, and Cécile Fougeron. 2002. "Realizations of Accentual Phrase in French Intonation." *Probus* 14 (1): 147–172.

Kaland, Constantijn. 2021. "Contour Clustering: A Field-Data-Driven Approach for Documenting and Analysing Prototypical F0 Contours." *Journal of the International Phonetic Association* 53 (1): 159–188.

Kaland, Constantijn. 2022. "Bending the String: Intonation Contour Length as a Correlate of Macro-Rhythm." In *Interspeech 2022*, 5233–5237. ISCA.

Kim, Sahyang, and Taehong Cho. 2009. "The Use of Phrase-Level Prosodic Information in Lexical Segmentation: Evidence from Word-Spotting Experiments in Korean." *The Journal of the Acoustical Society of America* 125 (5): 3373.

Kohler, Klaus J. 2009. "Rhythm in Speech and Language." *Phonetica* 66 (1–2): 29–45.

Ladd, D. Robert. 2008. *Intonational Phonology*, 2nd ed. Cambridge: Cambridge University Press.

Low, Ee Ling, Esther Grabe, and Francis Nolan. 2000. "Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English." *Language and Speech* 43 (4): 377–401.

Mennen, Ineke, Felix Schaeffler, and Gerard Docherty. 2012. "Cross-Language Differences in Fundamental Frequency Range: A Comparison of English and German." *The Journal of the Acoustical Society of America* 131 (3): 2249–2260.

Neal, Radford M. 2012. "MCMC Using Hamiltonian Dynamics."

Pierrehumbert, Janet. 1979. "The Perception of Fundamental Frequency Declination." *The Journal of the Acoustical Society of America* 66 (2): 363–369.

Pike, Kenneth L. 1945. "The Intonation of American English." Ann Arbor: University of Michigan Press.

Polyanskaya, Leona, Maria Grazia Busà, and Mikhail Ordin. 2020. "Capturing Cross-Linguistic Differences in Macro-Rhythm: The Case of Italian and English." *Language and Speech* 63 (2): 242–263.

Prechtel, Christine. 2019. "Quantifying Macro-Rhythm in English and Spanish." In *Proceedings of the 19th International Congress of Phonetic Sciences*, edited by Sasha Calhoun, Paola Escudero, and Marija Tabain, 2896–2900. Melbourne: Australasian Speech Science and Technology Association.

Prechtel, Christine. 2020a. "Quantifying Macro-Rhythm in English and Spanish: A Comparison of Tonal Rhythm Strength." MA Thesis, UCLA.

Prechtel, Christine. 2020b. "Macro-Rhythm in English and Spanish: Evidence from Radio Newscaster Speech." In *Proceedings of Speech Prosody 2020*, 675–679. ISCA.

R Core Team. 2020. "R: The R Project for Statistical Computing." https://www.r-project.org/

R Studio Team. 2020. "RStudio: Integrated Development for R." RStudio, Inc. https://www.rstudio.com/

Ramus, Franck, Marina Nespor, and Jacques Mehler. 2000. "Correlates of Linguistic Rhythm in the Speech Signal." *Cognition* 75 (1): AD3–30.

Rathcke, Tamara V., and Rachel H. Smith. 2015. "Speech Timing and Linguistic Rhythm: On the Acoustic Bases of Rhythm Typologies." *The Journal of the Acoustical Society of America* 137 (5): 2834–2845.

Salesky, Elizabeth, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. "Multilingual TEDx Corpus for Speech Recognition and Translation." In *Interspeech 2021*, 3655–3659.

Silverman, Kim, Mary Beckman, John Pitrelli, Mori Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. "ToBI: A Standard for Labeling English Prosody." In *Proceedings of the Second International Conference on Spoken Language Processing*, 867–870.

TED Conferences LLC. 1984. "TED: Ideas Worth Spreading." https://www.ted.com/

Van Heuven, Vincent J. 2018. "Acoustic Correlates and Perceptual Cues of Word and Sentence Stress: Towards a Cross-Linguistic Perspective." In *The Study of Word Stress and Accent*, edited by Rob Goedemans, Jeffrey Heinz, and H. Van der Hulst, 1st ed., 15–59. Cambridge: Cambridge University Press.

Vigário, Marina, and Sónia Frota. 2003. "The Intonation of Standard and Northern European Portuguese: A Comparative Intonational Phonology Approach." *Journal of Portuguese Linguistics* 2 (2): 115.

Wong, Wai Yi P., Marjorie K.M. Chan, and Mary E. Beckman. 2005. "An Autosegmental-Metrical Analysis and Prosodic Annotation Conventions for Cantonese." In *Prosodic Typology*, edited by Sun-Ah Jun, 271–299. Oxford: Oxford University Press.

Woodrow, Herbert. 1951. "Time Perception." In *Handbook of Experimental Psychology*, edited by S. Stevens, 1224–1236. Oxford: Wiley.

Appendix I. Transcripts of "The North Wind and the Sun"

English: The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off, should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew, the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

Greek: Ο βοριάς κι ο ήλιος μάλωναν για το ποιος απ' τους δυο είναι ο δυνατότερος, όταν έτυχε να περάσει από μπροστά τους ένας ταξιδιώτης που φορούσε κάπα. Όταν τον είδαν, ο βοριάς κι ο ήλιος συμφώνησαν ότι όποιος έκανε τον ταξιδιώτη να βγάλει την κάπα του θα θεωρούνταν ο πιο δυνατός. Ο βοριάς άρχισε τότε να φυσάει με μανία, αλλά όσο περισσότερο φυσούσε τόσο περισσότερο τυλιγόταν με την κάπα του ο ταξιδιώτης, ώσπου ο βοριάς κουράστηκε και σταμάτησε να φυσάει. Τότε ο ήλιος άρχισε με τη σειρά του να λάμπει δυνατά και γρήγορα ο ταξιδιώτης ζεστάθηκε κι έβγαλε την κάπα του. Έτσι ο βοριάς αναγκάστηκε να παραδεχτεί ότι ο ήλιος είναι πιο δυνατός απ' αυτόν.

German: Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges daherkam. Sie wurden einig, daß derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen. Der Nordwind blies mit aller Macht, aber je mehr er blies, desto fester hüllte sich der Wanderer in seinen Mantel ein. Endlich gab der Nordwind den Kampf auf. Nur erwärmte die Sonne die Luft mit ihren freundlichen Strahlen, und schon nach einigen Augenblicken zog der Wanderer seinen Mantel aus. Da mußte der Nordwind zugeben, daß die Sonne von ihnen beiden der Stärkere wäre.

European Portuguese: O vento norte e o sol discutiam sobre qual das duas era o mais forte, maas quando sucedeu passar um viajante envolto num manto. Ao vê-lo, põem-se de acordo em como aquele que primeiro conseguisse obrigar o viajante a tirar a capa seria considerado o mais forte. O vento norte começou a soprar com muita fúria, mas quanto mais soprava, mais o viajante se embrulhava no seu manto, até que o vento norte desistiu. O sol brilhou então com todo o seu esplendor, e imediatamente o viajou tirou o seu manto. E assim o vento norte teve de reconhecer a superioridade do sol.

French: La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort. Quand ils ont vu un voyageur qui s'avançait, enveloppé dans son manteau, ils sont tombés d'accord que celui qui arriverait le premier à le lui faire ôter serait regardé comme le plus fort. Alors, la bise s'est mise à souffler de toutes ses forces, mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui. Finalement, elle renonça à le lui faire ôter. Alors le soleil commença à briller et au bout d'un moment le voyageur, réchauffé, ôta son manteau. Ainsi, la bise dût reconnaître que le soleil était le plus fort.

Seoul Korean: 북풍과 햇님이 서로 힘이 더 세다고 다투고 있을 때, 한 내그네가 따뜻한 외투를 입고 걸어왔습니다. 그들은 누구든지 나그네의 외투를 먼저 벗기는 이가 힘이 더 세다고 하기로 결정했습니다. 북풍은 힘껏 불었으나, 불면 불수록 나그네는 외투를 단단히 여몄습니다. 그때에 햇님이 뜨거운 햇빛을 가만히 내려 쬐니, 나그네는 외투를 얼른 벗었습니다. 이리하여 북풍은 해님이 둘중에 힘이 더 세다고 인정하지 않을 수 없었습니다.

Appendix II. List of Korean TED Talks, including Youtube video ID, speaker, title and duration

Video ID	Speaker	Title	Duration
_1qZW7dL6TY	Chibok Song	Challenging the status quo by replacing the old constraints	18:45
_MzlPMY16Hw	Seungwoo Oh	New paradigm of technology and possibility on clothing	22:56
12K7MqoMpYk	Jinyong Kim	Counter the Corona Virus, but Prepare for the Return of the Corona Virus	22:32
4qCZ15jTZWc	Ho-Jeong Kang	CO_2 and wetland	18:13
aWPBoxeM8UA	Geon Hyeok Go	Sustainable DoReMi	17:31
XaE8qc5x5aY	Yeonho Oh	Every citizen is a journalist	17:16
gsxvPEFhXmE	Jaehyung Lee	The Path to Finding New Alternative Energy	16:52
gwg56V-GEMQ	Kyung-Mook Kim	Multiple identity diaspora	10:09
i2ye7MXvOcg	Jongbeom Lee	A Storyteller Explores Webtoons, Representing Korea	18:38
WRd7fHY3enA	KeeChang Kim	Inconvenient truth of Korean web	17:44
tmvodGy6RSM	Suk-jong Jang	3 epic years of Korean street fashion magazine	13:33
lGU_mqIdCAE	Myung-Sook Seo	Jeju Olleh	28:07
LnGadr1MDs8	Yeo-joon Yoon	Statecraft, the art of conducting public affairs	19:48
Mwo47MAsMtE	SanHa Kim	Gibbons, long-armed apes	20:02
MXvaF1nTkaY	Jungkyu Shin	Open source, not difficult	12:25
NahQP4v2QGo	Sang Min Whang	Coolguy and nobobody	22:16
NqNOcPFgmKo	Inhyuk Song	Organizer Creativity comes from 'between us'	16:28
o3_oYc38Uto	Yeon-jung Ahn	Making the invisible visible	15:16
RrYq5KRorSI	Doojin Hwang	My version of globalization	22:45
k1J7BP8fI8g	Min Ho Nam	Scientific approach of Korean medicine	17:02
SP67ty4SXQA	Doojin Hwang	Starting from where I am	17:50

Appendix II. *(continued)*

Video ID	Speaker	Title	Duration
t7toEV41rls	Geonhyeok Go	Indie, in itself should be the medium and the door	19:25
D35qys8YZpo	Jaram Lee	Telling the new story in the old form	21:24
TMxCKPNpfjY	Juhwan Kim	The Universe is a Romance	23:00
un4qbATrmx8	Kim Seung-beom	More humane medical experience	18:22
wcuRTEhhBao	Jaejun Lee	Designing fair Housing	19:30
zn8fTjUn_u8	YongSoon Park	Korean traditional hawk hunting	17:33
aja-8sYjNaY	Young-ha Kim	Be an artist, right now	18:09

CHAPTER 17

The phonetics and phonological behaviour of vowel length in Meto

Kirsten Culhane & Owen Edwards
University of Canterbury

This paper examines the acoustic cues and phonological behaviour of vowel length in Meto (Austronesian, Timor), which is described as having a distinction between "double" (/aa/ → [aː]) and "single" (/a/ → [a]) vowels. We examine acoustic evidence for this distinction, finding that putative double vowels are ~40% longer than single vowels. There is also some evidence of following consonant duration and vowel quality differences between single and double vowels.

We then examine the phonological behaviour of the putative double vowels, showing that their phonotactic patterning and (morpho)phonological behaviour is identical to unambiguous vowel sequences. This study highlights how field data can be used in phonetic and phonological research, and the importance of analysis in language documentation.

Keywords: vowel length, double vowels, phonological analysis, Timor

1. Introduction

Meto, an Austronesian language/dialect cluster of Timor, has been described by some analysts as having a phonological length distinction between single vowels, such as /a/, and double vowels, such as /aa/ (Edwards 2018, 2020; Culhane 2018, 2024). Examples illustrating the vowel length distinction are given in Table 1 (note that in Meto there are very few minimal pairs for vowel length.) However, this distinction is not recognised to the same extent or in the same environments by other analysts (Steinhauer 1996a, 1996b; Mooney 2022).

This paper examines both the acoustic cues and phonological behaviour of vowel length in Meto, addressing the following research questions:

https://doi.org/10.1075/slcs.240.17cul
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

Table 1. Kotos Amarasi short and long vowels

/V/ → [V]		Gloss	/VV/ → [V:]		Gloss
/asik/	[ʔasik]	'flea'	/na-siik/	[naˈsiːk]	's/he seizes'
/ike/	[ʔɪkɛ]	'spindle'	/bifee/	[bɪˈfɛː]	'woman'
/utan/	[ʔʊt̪an]	'vegetables'	/u-taan/	[ʔʊˈt̪aːn]	'I ask'
/knafo/	[ˈknafɔ]	'mouse, rat'	/na-foo/	[naˈfɔː]	'stinks'
/raku/	[ˈrakʊ]	'tuber'	/na-kuu/	[naˈkʊː]	's/he accuses'

1. Is there acoustic evidence for analysing Meto as having a vowel length distinction?
2. Is a single segment /V:/ or two segment /$V_{\alpha}V_{\alpha}$/ analysis better supported?

We begin the paper in Section 2 with an overview of the acoustic cues of vowel length and its phonological patterning cross-linguistically. This is followed in Section 3 by background on the language and data used in this paper. We then undertake an acoustic analysis of vowel length in one variety of Meto in Section 4, showing that putative double vowels are distinguished from single vowels by duration, and to some extent the duration of following consonants, and vowel quality. This is followed in Section 5 by an examination of how vowel length in Meto is best analysed phonologically. We show that evidence from phonotactics and morphophonology supports an analysis of Meto has having sequences of two identical vowels. This analysis is also supported by analytical parsimony. We conclude in Section 6 with the implications of our findings.

2. Background

2.1 Acoustic cues of vowel length

Many languages are analysed as having a phonological distinction between vowels of different perceived lengths. Most commonly, languages are reported to have a two-way length contrast between long and short vowels, while three-way length distinctions are rare cross-linguistically (Remijsen & Gilley 2008; Odden 2011).

The most salient cue that distinguishes long and short vowels cross-linguistically is vowel duration. However, both the raw duration values of long and short vowels, and the proportional difference between long and short vowels, varies considerably cross-linguistically (see Jepson 2025 for an overview). The duration of vowels is also affected by numerous other factors other than phonological vowel length, such as stress (Summers 1987; van Santen 1992), voicing of the following

consonant (Chen 1970; Kluender et al. 1988), syllable structure (Maddieson 1985), prosodic position and speech rate (Paschen et al. 2022), as well as vowel quality, as low vowels are typically longer than high vowels, all else being equal (Lehiste 1970; Keating 1985).

In addition to vowel duration, numerous other cues of vowel length have been identified. In many languages, vowel length is correlated with quality differences; where long vowels are often more peripheral in the vowel space and short vowels are more central in the vowel space (Lehiste 1970; Abramson & Reo 1990). Consonant length is also attested as a secondary cue of vowel length. Typically, there is an inverse relationship between consonant duration and phonological vowel length, whereby consonants are longer after short vowels, and shorter after long vowels. Differences in the duration of consonants after long and short vowels are observed both for codas and for onsets in the syllable immediately following a given vowel (see Jepson 2025 for an overview).

Additionally, numerous perception studies have found that dynamic f0 (changes in fundamental frequency over time) plays a role in whether listeners perceive vowels as short or long (Lehiste 1976; Lehnert-LeHouillier 2007; Hui & Arai 2019). However, there do not seem to have been any production studies demonstrating that dynamic f0 is an acoustic correlate of vowel length and as a result it is not investigated in the current study.

2.2 The phonological patterning of vowel length

There are two main ways vowel length distinctions can be analysed phonologically. The first option is to propose that long vowels have the status of single segments /V:/, which we refer to as a *segmental long vowel* analysis. If a language is analysed as having segmental long vowels, they are typically considered to be part of the vowel inventory.

The second option is to propose that they are sequences of two identical vowels, i.e. /$V_{\alpha}V_{\alpha}$/. We refer to this as a *double vowel* analysis. In such cases, long vowels are not analysed as part of the vowel inventory.

It is also worth noting that sometimes long vowels behave like both single segments and double vowels in one language. For example, in Markweeta (a.k.a. Endo, Nilotic, Kenya), segmental long vowels occur in roots. Evidence that they have the status of segments comes from reduplication; with long and short vowels reduplicated in the same way, see Zwarts (2004: 45). On the other hand, double vowels arise in Markweeta at morpheme boundaries (Table 6).

A range of criteria can be used to distinguish long single segments from sequences of identical segments. These are listed in (1).

(1) a. integrity (Kenstowicz & Pyle 1973; Hayes 1986a, 1986b)
b. inalterability (Hayes 1986a, 1986b)
c. phonotactic distribution (Trubetzkoy 1939; Harvey et al. 2019; Culhane 2022, 2024)
d. (morpho)phonological behaviour, especially morphological concatenation (Hayes 1986b; Odden 2011; Dmitrieva 2012; Culhane 2024)
e. analytical parsimony (Kehrein 2002; Harvey et al. 2019; Culhane 2022, 2024)

Integrity (1a) refers to the notion that if something is a single segment, it will not undergo processes which split up segment sequences, such as epenthesis or metathesis (Kenstowicz & Pyle 1973; Hayes 1986a, 1986b). On the other hand, if a process splits up long vowels, this provides evidence that they are better analysed as a sequence of identical vowels than a single segment. For instance, in Fur (Furan, Sudan), long vowels are split up by metathesis, as for other sequences of vowels. Examples are given in Table 2.

Table 2. Fur metathesis (Jakobi 1990: 46)

3SG.PF	1SG.PF	Translation
/ti**ir**e/	/**ir**e/	'meet, join'
/t**uu**mò/	/**ùtu**mò/	'build'
/**òò**ŋ/	/**o**ŋ**o**/	'sit'
/**ie**/	/**ĭ**re/	'snatch'
/ti**ò**/	/**i**to/	'wring, strain'

Inalterability (1b) refers to the prediction that the phonological behaviour of single long segments will be different to that of two short segments of the same kind. For example, in Daakaka (Austronesian, Vanuatu) words with long vowels are reduplicated differently from words with vowel sequences. When words with long vowels undergo partial reduplication of the first syllable, the long vowel is fully copied, e.g., /ja:~ja:/ 'REDUPLICATION~hurt' or /ku:~ku:/ 'REDUPLICATION~move'. On the other hand, for words with an initial CVV, only the first vowel is copied, for example /ku~kuɔ/ 'REDUPLICATION~run', /vɛ~vɛɔp/ 'REDUPLICATION~be.long' and /ga~gao/ 'REDUPLICATION~ dry' (von Prince 2015: 31).

An example of a language where vowel length demonstrates no evidence of being "inalterable" is Finnish (Uralic, Northern Europe). In Finnish, putative double vowels have the same phonological behaviour as vowel sequences. For example, the first vowel of a vowel sequence is deleted when the plural suffix /-i/

is attached. This process affects both sequences of different vowels and putative double vowels in the same way. Examples are given in Table 3.

Table 3. Finnish suffixation (Odden 2011: 4)

Singular	Plural	
/s**uo**-ta/	/so-i-ta/	'swamp'
/ty**ø**-ta/	/t**ø**-i-ta/	'work'
/t**ie**-tæ/	/te-i-tæ/	'road'
/m**aa**-ta/	/ma-i-ta/	'earth'
/s**yy**-tæ/	/sy-i-tæ/	'reason'

Phonotactic distribution arguments (1c) involve comparing the distribution of the phenomenon in question to that of single segments or sequences (Harvey et al. 2019; Culhane 2022, 2024). This is based on the idea that phenomena which can be analysed as either a single segment or sequence (in this case, vowel length) should be interpreted as single segments if they occur in environments where sequences do not occur (Trubetzkoy 1939, Rule IV). On the other hand, if the phenomenon in question (e.g., long vowels) has the same phonotactic distribution as sequences (of vowels), this provides support for analysing it as a sequence. An example of such a language where this is the case is Belep (Austronesian, New Caledonia). In Belep, both long vowels and vowel sequences are only attested word-finally. Examples are given in Table 4, with long vowels represented as double vowels.

Table 4. Phonotactic distribution of Belep /VV/ (McCracken 2019: 68ff)

Word shape	Example	Gloss
CV**V**	/n**ao**/	'to sing'
	/^{n}d**ee**/	'blacksaddled coral grouper'
CV**V**C	/w**au**p/	'to be toothless'
	/k**aa**t/	'butterfly fish'
CVC**V**V	/^{m}bol**ao**/	'banana tree'
	/cẽn**ee**/	'to swallow'
CVCVC**V**V	/^{m}b^{w}alaw**io**/	'place name'
	/karav**aa**/	'pirogue'

A language in which long vowels have a different distribution to vowel sequences — lending support for *not* analysing them as sequences of identical vowels — is Rukai (Austronesian, Taiwan). In Rukai, long vowels occur almost exclusively in penultimate syllables, while vowel sequences occur in other word positions. Examples are given in Table 5.

Table 5. Phonotactic distribution of Rukai /V:/ (Chen 2006: 214ff)

Example	Gloss
/**a:**bu/	'ashes'
/**i:**pi/	'blow'
/kə**:**tə/	'cut'
/aða**:**ðamə/	'bird'
/maθ**o:**lo/	'cough'
/is**iu**/	'hair'
/iak**ai**/	'to exist'
/θal**ia**gumu/	'ipomea'

(Morpho)phonological behaviour can also be relevant when deciding on the best analysis of vowel length (1d). If long vowels arise due to morphological concatenation (i.e. /V_α-V_α/ → [V:]) this can also support a double vowel analysis, with one vowel coming from each morpheme. This kind of evidence is also often used to support a two-segment analysis of long consonants (e.g. Hayes 1986b; Dmitrieva 2012) but has also been used to support a double vowel analysis of long vowels (Odden 2011; Culhane 2024). An example of a language where this occurs is Markweeta. In Markweeta long vowels arise when a vowel-final morpheme is followed by a vocalic morpheme. Examples are given in Table 6.

Table 6. Markweeta double vowels (Zwarts 2004: 24)

/wę̱ę̱r**i̱**-**i**/	son-DEF	[wę̱**:**ri̱**:**]	'the son'
/ma-**a**-nke̱t/	NEG-1SG-know	[mą̱**:**nké̱t]	'I don't know'
/k**i**-**i**-cam/	DISTANT.PAST-2SG-accept	[kí**:**càm]	'you have accepted'

Analytical parsimony can also be used as a basis for analysing a language as having segmental long vowels or double vowels (1e). Analytical parsimony arguments typically involve comparing the size of the segment inventory and the phonotactic restrictions required to account for the distribution of the phenomenon in question under either a single segment or sequence analysis. Under a single

segment (e.g. segmental long vowel) analysis, the segmental inventory is larger, and may mean that fewer restrictions are required to account for the phonotactic distribution of the phenomenon in question. On the other hand, under a sequence (e.g. double vowel) analysis, the segmental inventory is smaller, but it may result in a more complex description of the permitted sequences and their phonotactic distribution (Kehrein 2002; Harvey et al. 2019; Culhane 2022).

For example, if a language attests many long vowels, as well as sequences of two vowels, it may be in in the interests of analytical parsimony to propose that they are sequences of segments, rather than proposing additional vowels in the inventory. An example of a language where this argument can be made is Māori (Austronesian, Aotearoa New Zealand). Māori is typically analysed as having a segmental inventory of 10 consonants /p, t, k, m, n, ŋ, w, f, ɾ, h/ and five vowels /i, e, a, o, u/, which can occur in any combination (Harlow 2007; Bauer 1993). This includes sequences of identical vowels. Analysing Māori as having segmental long vowels would double the size of the vowel inventory, and does not yield a simpler account of the phonotactics, as other vowel sequences are otherwise attested.

On the other hand, if only a few long vowels are attested, and/or no other vowel sequences are attested, it is in the interests of analytical parsimony to increase the segmental inventory size rather than to propose that a very small number of vowel sequences are permitted. Djambarrpuyŋu (Pama-Nyungan, Australia) is an example of a language for which this argument can be made. Djambarrpuyŋu attests vowel length distinctions for all three of its vowels, e.g. /k**u**rum/ 'soft' vs. /k**u:**rum/ 'to be high', /w**a**ŋa/ 'talk' vs. /w**a:**ŋa/ 'home', /wiːʈi/ 'wallaby' vs. /wiʈːic/ 'olive python' (Jepson 2019: 2; Wilkinson 1991: 44). Djambarrpuyŋu does not attest vowel sequences, and so analysing the long vowels as sequences would require proposing that sequences of identical vowels i.e. /ii aa uu/ are attested, but no other vowel sequences.

It is worth stressing that the acoustic cues of vowel length cannot be transparently related to their phonological patterning. We may expect, for example, that if a language attests sequences of two identical vowels, they will be approximately twice the duration of monophthongs. However, this is not the case. There are languages where long vowels are approximately twice as long as short vowels but pattern phonologically as single segments (e.g. Djambarrpuyŋu [Jepson 2019, 2025]). There are also languages where putative double vowels are not double the length of single vowels. An example is Qaqet (Baining, Papua New Guinea), where long vowels almost always occur at morpheme boundaries, (e.g. /a-a/, Hellwig 2019: 41), but the duration of these putative double vowels is only approximately 1.4–1.6 times the duration of short vowels (Tabain & Hellwig 2023: 1085).

3. Language background

Meto (a.k.a. Uab Meto, Dawan[ese], Timorese, Atoni) is an Austronesian language/dialect cluster spoken on the western part of the island of Timor; both in the East Timorese enclave of Oecusse, as well as in the Indonesian province of Nusa Tenggara Timur.

Meto speakers think of their speech as a single language, but differences in function words, lexicon, and morphophonemic processes accumulate between different varieties of Meto to such an extent that mutual intelligibility is often blocked. Meto speakers recognise more than a dozen named varieties of Meto, which themselves have named dialects, with further differences found between different villages and hamlets of a single dialect. A map showing the varieties of Meto is given in Figure 1 for reference, as we present data from a number of different varieties in this paper.

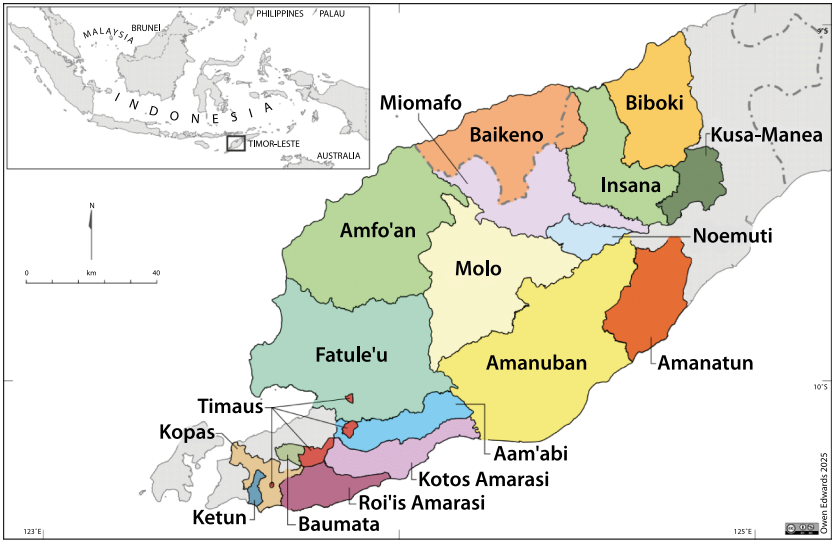


Figure 1. Varieties of Meto

All Meto varieties have five contrastive vowels /i e a o u/. The mid vowels display variation between mid-low [ɛ ɔ] and mid-high [e o]. Sequences of two different vowels are widely attested, and segmental diphthongs are not attested (Culhane 2024). All (known) Meto varieties have 11 core consonants /p t k ʔ b f s h m n/ and either /r/ or /l/. Some varieties (e.g. Timaus and Kusa-Manea) have both /r/ and /l/ and thus have 12 core consonants. In addition to these consonants, most varieties of Meto also have a voiced palatal obstruent /ʤ/ and a voiced velar

obstruent, either /gʷ/ or /g/, though some lack the voiced velar obstruent. The voiced palatal and velar obstruents are restricted in distribution, and mostly only occur under processes of consonant insertion. Amanuban and Amanatun lack these obstruents but have glides /j/ and /w/ in comparable environments.

All described varieties of Meto have been reported to have stress falling on the penultimate syllable of the word. Culhane (2024) demonstrates that there is acoustic evidence for stress in Amfo'an, with significant differences in spectral tilt, vowel quality and duration being found between stressed and unstressed syllables.

All Meto data in this paper was collected by the authors, with the exception of some Baikeno data which was collected and provided by Charles E. Grimes. Our most comprehensive data is from Kotos Amarasi (Nekmese' village, Koro'oto hamlet), Roi'is Amarasi (Buraen and Tunbaun villages), and Nai'bais Amfo'an (Lelogama village). However, we have collected data on all varieties of Meto (except Biboki) and find that the double vowel analysis proposed in this paper is appropriate for all varieties. All Meto data is labelled in this paper according to the variety from which it comes and, when relevant, the village in which it was collected.

4. Acoustic cues of vowel length in Meto

This section presents a brief study of the acoustic cues of vowel length in Meto, with a focus on data from the Amfo'an variety. It follows a similar methodology to that of Jepson (2025). The following subsections outline the data used, how it was processed, and the statistical analysis. Section 4.4 presents the results. The full data set, code and statistical analysis results can be found at https://osf.io/jahwm/.

4.1 Data selection and annotation

The data used here is taken from a larger database of Amfo'an data collected by the first author, which is archived in PARADISEC (Culhane 2017). These are audio recordings made in 2017 using a Zoom H1 recorder. The data comprises spontaneous first-person narratives of varying lengths, and come from seven native speakers of Amfo'an. Two of these speakers are women, and five are men. They were aged between 17 and 71 years old at the time of recording. The data was segmented in Praat (Boersma & Weenink 2023) by the first author. Recordings in the database are transcribed and segmented into intonation units, as well as words, syllables and segments.

The data used in this study comprises all instances of "single" /i e a o u/ and "double" vowels /ii ee aa oo uu/ in the database, excluding loanwords from Malay and vowels which occurred in a sequence with another vowel (e.g. /ai/, /oe/ etc). The data used here also includes all consonants following single and double vowels, either in a coda (i.e. CVV**C**, CV**C**) or in the onset of a following syllable (i.e. CVV.**C**V, CV.**C**V). The vowels and consonants we examine in our study were annotated for what kind of syllable they occur in (open or closed), if the syllable is stressed (stressed or unstressed) and the phrase position of the word a given syllable occurs in, making a five-way distinction between first, second, medial, prefinal and final positions.

4.2 Measurements, outlier removal and normalisation

We examine three possible cues of vowel length: vowel duration, consonant duration, and vowel quality (formants). The duration of all vowels in the data set and the following consonants were extracted using a Praat script from Kaland (2021). For vowel duration, observations 2.5 standard deviations above or below the mean for each vowel quality were excluded. For consonant duration, observations 2.5 standard deviations above or below the mean for a given consonant were excluded. Both vowel and consonant duration were z-scored. Z-scoring is a way of standardising data, whereby raw scores are converted into a common scale with a mean of 0. The converted z-score represents how many standard deviations away from the mean a given data point is. Following outlier removal, the final duration data set contained 2093 vowel tokens and 1707 consonant tokens, which are summarised in Table 7.

For each vowel, F1 and F2 measurements were taken from the voiced portions of syllables where formants demonstrate stable values. Both the automatic labelling of subintervals and formant measurements were performed using scripts from Kaland (2021). Observations 2.5 standard deviations above or below the mean for each formant and for each vowel quality were excluded. Vowel formant values were normalised using the Lobanov-2.0 method (Brand et al. 2021) using the nzilbb.vowels() package (Wilson Black & Brand 2022) in R. Vowel normalisation helps to make formant data comparable across individual speakers, helping to control for physiological differences such as vocal tract length and vowel space size. In the Lobanov-2.0 method, the mean value of each formant for each vowel is calculated. Then, a mean of means (the mean value calculated across the individual vowel means) is calculated. This mean of means is subtracted from each raw formant value. Then, this value is divided by the standard deviation of the mean of means, giving the normalised formant value. This process results in the centre of the vowel space being (0,0) for all speakers, and is done so that formant values for

Table 7. Data overview

a. Vowels

Vowel quality	Single	Double
i	308	60
e	367	135
a	722	84
o	206	39
u	142	36

b. Consonants

Consonant type		Previous vowel	
		Single	Double
Coda	Sonorant	356	161
	Obstruent	322	83
VCV	Sonorant	369	33
	Obstruent	343	40

different speakers cover approximately the same space. Lobanov-2.0 is used here because it is well-suited for imbalanced data sets (e.g. where there are not equal numbers of each vowel). See Brand et al. (2021) for details.

4.3 Statistical analysis

We perform linear mixed-effects regression analyses using the lme4 package (Bates et al. 2015) in R (R Core Team 2024). A separate analysis is performed for each of the following: z-scored vowel duration, z-scored consonant duration, F1 (Lobanov-2.0 normalized) and F2 (Lobanov- 2.0 normalized).

For vowel duration, we examine the effects of vowel type (single vs. double) and the interaction of vowel type with vowel quality. To ensure, as much as possible, that possible environment effects on vowel type are controlled for, we include the following control predictors: phrase position, syllable type, and syllables per second. Syllables per second was calculated per intonation unit and is included as a local measure of speech rate. Stress is included as a control predictor in the vowel duration model because duration has previously been found to be an acoustic correlate of stress in Amfo'an (see Culhane [2024] for details).

For consonant duration, we examine the effects of previous vowel type (single vs double), and the interaction of previous vowel type with consonant type

(obstruent or sonorant) and consonant position (coda or intervocalic). Stress, phrase position and syllables per second are included as control predictors.

For F1 and F2, we examine the interaction of vowel type and vowel quality, with control predictors for following consonant place (coronal vs. dorsal vs. other) and stress. Stress is included in these models as vowel quality has also been found to be a correlate of stress in Amfo'an (Culhane 2024). All models include random intercepts for speaker and word. In all models, vowel type is a sum-coded binary factor,[1] where vowel type = double is positive. Stress is a sum-coded binary factor, where stress = TRUE is positive. Vowel is a sum-coded factor with five levels, where each vowel is compared to the overall mean. Phrase position is also a sum-coded factor with five levels. syllable type, consonant type and consonant position are sum-coded binary factors, whereby syllable type = closed, consonant type = sonorant and consonant position = coda are positive. To aid the interpretation of significant effects, we conduct model comparisons using the anova() function. For significant interactions, we conduct EMM tests and pairwise comparisons using the emmeans package (Lenth 2024).

4.4 Results

4.4.1 *Vowel duration*

Table 8 provides a summary of the mean duration values in milliseconds and standard deviation (in brackets) for each vowel quality and vowel type (i.e. "single" or "double" vowels). It also gives the ratio of means (single vs double).

We can see from Table 8 that the raw duration values of double vowels are greater than for single vowels. This can also be seen in Figure 2. We can also see in Figure 2 that duration values are more variable for double vowels than single vowels.[2] The ratio of double vs single vowels varies between vowel qualities, but averaging across the different vowel qualities, putative double vowels are overall 40% longer than single vowels.

Our analysis reveals significant interactions between vowel type and vowel quality. Model comparison confirms that there is a significant overall effect of vowel type ($\chi 2 = 74.61$, $df = 5$, $p < 0.001$). To investigate differences between double

1. When categorical variables are sum-coded, the different levels of the variable are coded so that their levels are compared to the grand mean, and there is no reference value. For example, when a categorical variable has two levels, they are given the values –1 and +1. The intercept in the model is the grand mean, and the coefficient for each level represents the difference between the mean for that level and the overall grand mean.

2. Note that in Figure 2 duration values are z-scored by phrase position to control for effects of final lengthening.

Table 8. Summary of mean duration values (ms), standard deviations (ms), and ratios for vowels

Vowel	Single	Double	Ratio (single:double)
i	105.96 (67.85)	174.03 (83.31)	1:1.64
e	102.50 (45.92)	125.57 (62.60)	1:1.23
a	96.92 (42.95)	140.50 (57.64)	1:1.45
o	124.26 (56.76)	170.75 (63.53)	1:1.37
u	102.18 (50.49)	144.05 (63.69)	1:1.41
overall	103.36 (68.08)	103.36 (51.70)	1:1.4

and single vowels for each vowel, we conduct an EMM test and pairwise comparisons. This allows us to examine the differences between, for example, single /i/ and double /ii/. This indicated that double vowels are significantly longer than single vowels for all vowel qualities. The results are given in Table 9. Note that in 9 the contrast is between single and double vowels, where negative estimates capture that single vowels are shorter than double vowels.

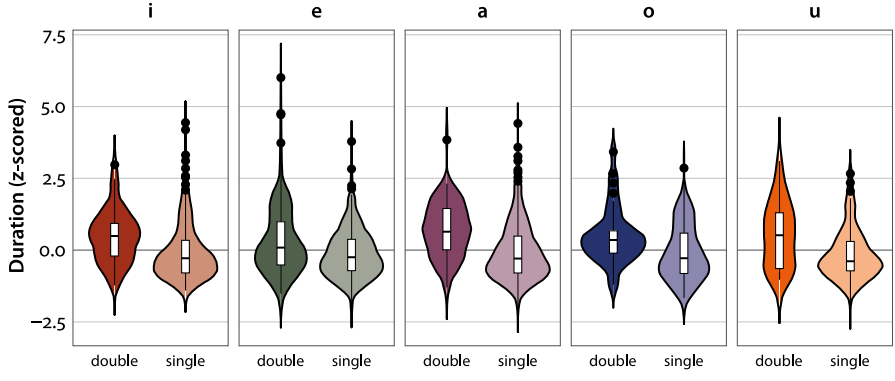


Figure 2. Duration values of single and double vowels

Table 9. Pairwise comparisons: Single vs double vowel, by vowel quality

Vowel	β	SE	df	*t*	*P*
i	−0.568	0.137	780	−4.143	<0.001
e	−0.288	0.086	1416	−3.357	<0.001
a	−0.605	0.098	1530	−6.136	<0.001
o	−0.597	0.144	1727	−4.156	<0.001
u	−0.369	0.152	1715	−2.430	=0.015

4.4.2 *Consonant duration*

Table 10 provides a summary of the mean duration values in milliseconds and standard deviation (in brackets) for consonants following single and double vowels. The data are separated by consonant position (coda vs. intervocalic) and consonant type (sonorant vs obstruent).

Table 10. Summary of mean duration values (ms), standard deviation (ms) for consonants

Consonant type		**Previous vowel**	
		Single	**Double**
Coda	Sonorant	92.26 (33.85)	109.99 (39.44)
	Obstruent	93.51 (56.97)	121.99 (79.35)
VCV	Sonorant	69.73 (18.71)	81.58 (38.36)
	Obstruent	85.21 (39.52)	80.03 (46.89)

We can see in Table 10 that the raw duration of consonants following single vowels is greater in one category (intervocalic obstruent) but not in any other category. Additionally, this difference is very small and the standard deviations are very large.

However, when the data is z-scored, we see instead that consonants are indeed greater in duration after single vowels than double vowels, in all categories. This is because z-scored data captures how far a data point is from the mean of a dataset. In our case, the individual values for single vowels are further from the overall mean than the individual values for double vowels. This further distance from the mean is then reflected in higher z-scored values. This can be seen in Figure 3.

Our analysis reveals a significant overall effect of interaction of vowel type (double vs single) with consonant type (obstruent.v.sonorant, $\beta = 0.072$, $SE = 0.033$, $t(1474.58) = 2.19$, $p = 0.029$). To investigate further, we conducted pairwise comparisons. This indicated that obstruents are significantly longer after single vowels than after double vowels (single – double: $\beta = 0.21$, $SE = 0.10$, $t(1485) = 2.14$, $p = 0.033$).

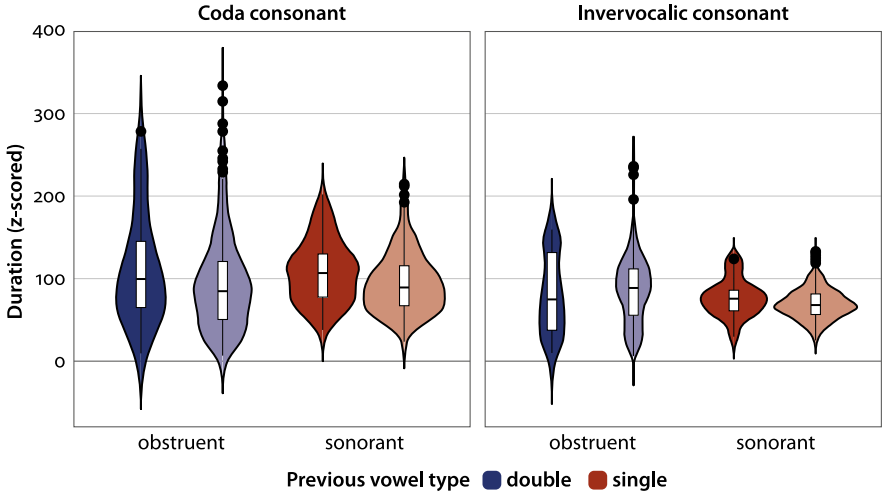


Figure 3. Duration values of consonants

4.4.3 *Vowel quality*

Figure 4 shows the normalised first and second formant values of single and double vowels in the data set. The level parameter for the ellipses in Figure 4 was set to 0.67, which corresponds to about one standard deviation from the mean. We can see in Figure 4 that double vowels appear to be more peripheral in the vowel space than single vowels.

Our analysis reveals that for F1, there is a significant interaction of vowel type and quality. To investigate further, we conduct an EMM test on the interaction of vowel type and quality, we find a significant difference between the F1 of single and double vowels for /a/ (single–double: $\beta = -0.694$, $SE = 0.117$, $t(1394) = -5.959$, $p < 0.001$). This entails that /aa/ has higher normalised F1 than /a/, corresponding to being lower in the vowel space.

For F2, there is also a significant interaction of vowel type and quality. An EMM test and pairwise comparison indicates that there is a significant difference in F2 between single /i/ and double /ii/ : (single-double: $\beta = -0.349$, $SE = 0.104$, $t(841) = -3.358$, $p = 0.0008$). Double /ii/ has a higher normalised F2 than /i/, meaning it is further forward in the vowel space than single /i/.

Pairwise comparison also indicates that that there is a significant difference in F2 between single /u/ and double /uu/: (single-double: $\beta = 0.303$, $SE = 0.112$, $t(1645) = 2.705$, $p = 0.0069$). This entails that double /uu/ has lower normalised F2 than /u/, meaning it is overall further back in the vowel space than single /u/.

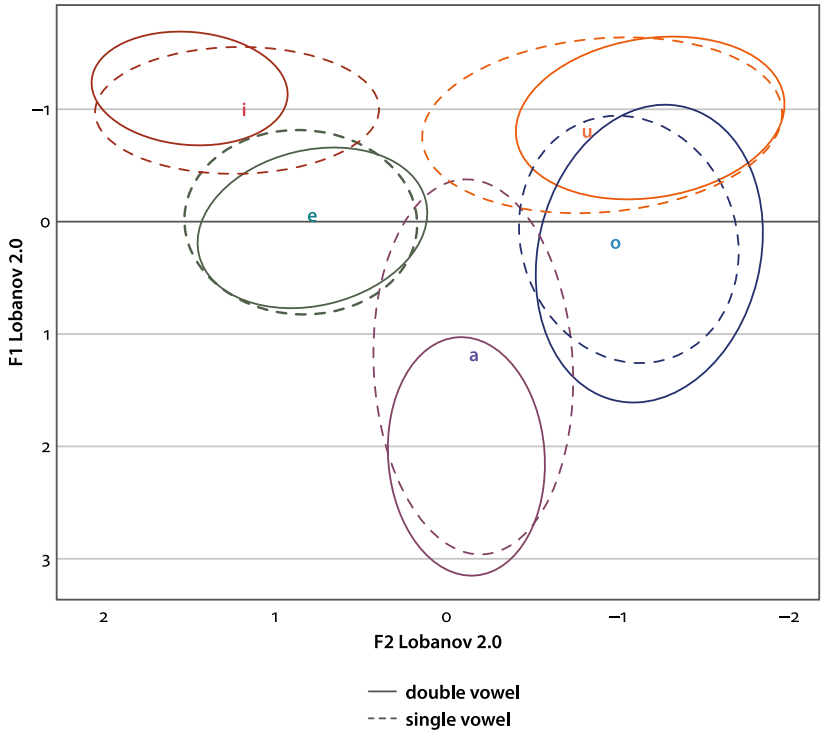


Figure 4. Normalised first and second formant values of single and double vowels in Amfo'an

4.4.4 *Summary*

This section has presented an acoustic study of the acoustic cues of vowel length in Amfo'an. It has shown that double vowels are statistically significantly longer than single vowels; approximately 40% longer. It has also found that consonant duration is a cue of vowel length to some extent,[3] whereby obstruents are shorter after double vowels than after single vowels. There is also some evidence that

3. A reviewer suggested that consonant length could be distinctive, with vowel length as a secondary cue. There are several reasons for not taking this analysis: (1) Not all long vowels are followed by a consonant (e.g. /bifee/ [bɪˈfɛː] 'woman'). (2) The putative double vowels are usually the result of historic consonant deletion (e.g. Proto-Malayo Polynesian *suqun > Meto /n-suu/ 'carry on head' (3) Cross-linguistically the relationship between vowel and consonant length is unclear for long consonants: in some cases, vowels before geminate consonants are shorter (e.g. Italian [Esposito & Di Benedetto 1999] Tashlhiyt Berber [Ridouane 2007], Bengali [Lahiri & Hankamer 1988]). In some cases they are longer (e.g. Japanese [Idemaru & Guion 2008]) and in some cases there is no significant difference in vowel duration (e.g. Turkish [Lahiri & Hankamer 1988]).

vowel quality may be associated with vowel length, but we only find significant effects for /i/, /a/, and /u/. These findings support an analysis of Meto as having a phonological vowel length distinction.

5. Phonological behaviour of vowel length in Meto

Having established that there is acoustic evidence for analysing Meto as having a phonological vowel length distinction, in this section, we examine the phonological patterning of putative double vowels, appealing to the criteria outlined in (1). In particular, we focus on their phonotactic distribution and morphophonological behaviour, finding that long vowels in Meto demonstrate identical behaviour to sequences of two different vowels.

5.1 Phonotactic distribution (criterion (1c))

In this section we discuss the phonotactic and metrical structures of Meto. Our discussion here is restricted to content/lexical words. Function/grammatical words have slightly different phonotactic structures (Edwards 2020: 145f). The initial part of our discussion also completely excludes all words which contain a putative double/long vowel.

Meto can be described as having disyllabic trochees which are maximally CCV́CVC and minimally CV́V. As mentioned above, Culhane (2024: 134–147) demonstrates that there is acoustic evidence for stress in Amfo'an from duration, spectral tilt, and to some extent, vowel quality. Culhane argues that a foot-based analysis of stress in Amfo'an is possible, but prefers appealing to the prosodic word. Under a foot-based analysis of stress (and leaving aside all putative double/long vowels), Meto can be said to have a weight-insensitive trochaic foot structure, with stress assigned to the penultimate syllable.

This weight-insensitive trochee must have an initial consonant, if none is specified a glottal stop is automatically inserted (Edwards 2017). The V-slot is filled by one of the five vowels. Initial and medial C-slots are filled by any consonant, and the final C-slot is filled by any one of the 12 core consonants (Section 3). Vowel sequences do not (obligatorily) diphthongise or otherwise form heavy syllables and stress is assigned to the penultimate syllable of such words — as it is for all words, e.g. Kotos Amarasi /kuan/ → ['kʊ.an] 'village'.

Glides, in those varieties of Meto with them (Amanuban and Amanutan), cannot occur in codas. No variety of Meto attests phonemic diphthongs or phonemic VG# sequences. The only diphthongs that arise are due to the (optional) coa-

lescence of two vowels, e.g. Kotos Amarasi /tei/ 'faeces' → ['t̪e.i] ~ [t̪ej] (Edwards 2020: 107).

With those considerations in mind, Table 11 summarises the root structures found in Kotos Amarasi, the variety of Meto for which the most electronic lexical data is available. Stress is indicated with an acute accent, putative secondary stress with a grave accent and foot boundaries are separated with a vertical bar |.[4]

Table 11. Kotos amarasi root structures

Syllables	Structure		
σσ	(C)CV́(C)V(C)	1,502	86%
σσσ	(ʔ)CV(C)|CV́(C)V(C)	156	9%
σσσσ	(ʔ)CV̀V(C)|CV́(C)V(C)	58	3%
σσσσ	(ʔ)CV̀CVC|CV́(C)V(C)	24	1%
		1,772	

The following facts summarise the root structures found in Meto:

1. Lexical roots in Meto are minimally disyllabic.
2. Consonant clusters only occur at the juncture of a foot and/or word initially,[5]
3. Vowel sequences do not occur across a foot boundary.
4. All feet must have an onset consonant.

While there are some slight differences between individual varieties of Meto in the kind of root structures found, these do not affect any of these generalisations. For example, while Kotos Amarasi does not allow CV̀CV|CV́CV(C) words, such words do occur in Amfo'an.

Vowel sequences have a restricted distribution, occurring only in the single/final foot of a root, or within the initial foot of a quadrisyllabic root. The possible locations of vowel sequences within Meto roots are summarised in Table 12a. The distribution of putative double/long vowels is shown in Table 12b. Surface long vowels have exactly the same distribution as vowel sequences, and occur in all positions in which vowel sequences occur.

4. Table 11 is based on a draft dictionary with 1,913 unique lexical roots. Ten roots with exceptional structures are excluded, as are the 131 roots with a putative double vowel, leaving 1,772 unique lexical roots. 32 roots with the structure (C)VVCV(C) are counted as disyllabic. See Edwards (2020: 106) for the evidence for this analysis.

5. Word-initial clusters for roots with more than two syllables are restricted to /ʔC/. Such roots are extremely rare with only eight instances in our current database.

Table 12. Root structures with [VV] and [V:]

(a) Roots with [VV] **Structure**	**(b) Roots with surface [V:]** **Surface structure**
(C)C<u>V́V</u>(C)(ʔ)CV(C)|C<u>V́V</u>(C)	(C)C<u>V́:</u>(C)
(ʔ)C<u>V̀V</u>(C)|C<u>V́V</u>(C)	(ʔ)CV(C)C<u>V́:</u>(C)
(ʔ)C<u>V̀V</u>(C)|CV́CV(C)	(ʔ)C<u>V̀:</u>(C)C<u>V́:</u>(C)
(ʔ)CV̀CVC|C<u>V́V</u>(C)	(ʔ)CV̀V(C)C<u>V́:</u>(C)
	(ʔ)C<u>V̀:</u>(C)CV́(C)V(C)
	(ʔ)CVCVCC<u>V́:</u>(C)

Putative double vowels have the same restricted occurrence within the Meto root as vowel sequences. The restrictions on the distribution of double vowels are poorly explained if they are analysed as segmental long vowels. On the other hand, if analysed as double vowels, the distribution of long vowels is identical to that of vowel sequences. This can be seen in Table 13 which shows examples of vowel sequences, including putative double vowels, in Kotos Amarasi.

If we wanted to maintain an analysis of Meto as having segmental long vowels, we could propose that the foot in Meto is weight-sensitive, with stress attracted to a final heavy syllable. This analysis is more complex than the proposed double vowel analysis, and would mean analysing Meto as having a vowel length distinction in final syllables, i.e. CVCV:C, but not in penultimate syllables, i.e. *CV:CVC. This is typologically unexpected; cross-linguistically there is a tendency to avoid length distinctions word-finally (Barnes 2002; Myers & Hansen 2007). While this analysis is possible, notably absent in Table 12b are disyllables with an initial long vowel, and final short vowel *(C)C<u>V́:</u>CV(C). Under a segmental long vowel analysis, this absence is difficult to account for. The lack of long vowels in penultimate syllables is also unexpected given the phonotactic patterns of the language more broadly. The penultimate syllable position allows a greater range of syllable structures than other syllable positions (most onset clusters occur in this position) and demonstrates various other characteristics of being the most prominent syllable in the word. For examine, in several varieties of Meto, the penultimate syllable triggers processes of vowel assimilation, with the final syllable assimilating to the quality of the penultimate syllable (Edwards 2020:181). It would thus be unexpected for Meto to lack a vowel length contrast in the most prominent syllable, while having a vowel length contrast in the final syllable of the word.

Furthermore, it is not the case that this penultimate syllable is always word initial. For trisyllabic and quadrisyllabic roots it is still the penultimate syllable (e.g. (ʔ)CV(C)|**CV́**(C)V(C) which is prominent and has these properties. Thus,

Table 13. Phonotactic distribution of vowel sequences in Kotos Amarasi

Putative word shape	Example		Gloss
CVV	/hau/	[ˈha.ʊ]	‘wood, tree’
	/nua/	[ˈnʊɐ]	‘two’
	/haa/	[hɐː]	‘four’
CVVC	/noah/	[ˈnɔ.ɐh]	‘coconut’
	/fauk/	[fɐwk]	‘how many’
	/huun/	[huːn]	‘grass’
CCVV	/ʔsao/	[ʔˌsa.ɔ]	‘crab’
	/krei/	[krei]	‘church, week’
	/ʔroo/	[ʔrɔː]	‘far’
CVCVV	/biʤae/	[biʤa.ɛ]	‘cow’
	/sekau/	[sɛˈkəw]	‘who’
	/bifee/	[biˈfɛː]	‘female, woman’
CVVCVCV(C)	/tainonus/	[t̪ajˈnonos]	‘earthquake’
	/paumakaʔ/	[ˌpɐwˈmakɐʔ]	‘near’
	/suufnenɛʔ/	[ˌsʊːfˈnɛnɛʔ]	‘k.o. tree snake’
CVVCVV(C)	/mainuan/	[majnʊ.ɐn]	‘open, opportunity’
	/naiʤeer/	[najˈʤɛːr]	‘ginger’
	/tootiu/	[t̪ɔːˈt̪i.ʊ]	‘k.o. bird’
CVVCCVV(C)	/riuksaen/	[riʊkˈsa.ɛn]	‘python’
	/aushee-f/	[ʔaʊsˈhɛːf]	‘spleen’
	/saanʔoo/	[saːnˈʔɔː]	‘stick insect’

this cannot be explained as being due to extraprosodicity, or, in OT-terms, to an effect of Positional Faithfulness (Beckman 1998).[6]

To summarise, the phonotactic restrictions on the distribution of putative double vowels within the Meto root would be poorly explained if they are analysed as segmental long vowels and would have a strange distribution given the trochaic foot-structure of Meto. However, no problems arise if they are analysed as double vowels. They have exactly the same distribution as other vowel sequences, and obey the same phonotactic restrictions as other vowel sequences.

6. The only feature of root structures that could be explained in these terms is the permissibility of word-initial /ʔC/ clusters in roots greater than a disyllable.

5.2 Morphophonological behaviour

5.2.1 *Evidence for a lack of integrity (criterion (1a)): Glottal stop infixation*

The first morphological process we consider is a process of glottal stop infixation whereby the final element of certain circumfixes has an infixal allomorph. The data in the following section comes from Kotos Amarasi. The circumfixes with an infixal allomorph are /ma-…-ʔ/ and /ʔ-…-ʔ/, which are both nominalisers. /ma-…-ʔ/ derives nouns with the semantics of describing a resulting state or of particular characterisation encoded by the root,[7] and /ʔ-…-ʔ/ typically derives nouns referring to tools. When these affixes attach to roots ending in CVCV(C) the final glottal stop occurs after the root. If the medial consonant is itself a glottal stop (CVʔV) the suffix has the allomorph /-k/. Examples are given in Table 14 from Kotos Amarasi. When these affixes occur on roots with final Va(C) the glottal stop also occurs as a suffix. Three examples are /fua-f/ 'fruit' → /ma-fua-ʔ/ 'fruit bearing' and /na-tua/'fill' → /ma-tua-ʔ/ 'filled'.

When these affixes occur on roots with any other vowel sequence, the glottal stop occurs as an infix between the final two vowels. The same pattern is found for words with a putative double vowel (other than /aa/). Examples are shown in Table 15.

Under an analysis of Meto as having double vowels, the occurrence of the glottal stop word medially in forms such as /ma-fe‹ʔ›e/ 'having a wife' is consistent with other roots with a vowel sequence, such as /ma-tu‹ʔ›i/ 'written'. Because they can be broken up by infixation, Meto double vowels demonstrate a lack of integrity (criterion (1a)).

Table 14. Nominalising /ma-…-ʔ/ and /ʔ-…-ʔ/, CVCV(C)# roots

Root gloss	Stem	Circumfix	Derivation	Gloss
'rock, stone'	/fatu/	+ /ma-…-ʔ/	→ /ma-fatu-ʔ/	'rocky, stony'
'wing'	/niniʔ/	+ /ma-…-ʔ/	→ /ma-nini-ʔ/	'winged'
'price'	/osa-f/	+ /ma-…-ʔ/	→ /ma-ʔosa-ʔ/	'expensive, valuable'
'hear'	/n-nena/	+ /ma-…-ʔ/	→ /ma-nena-ʔ/	'heard'
'call, name'	/n-teka/	+ /ma-…-ʔ/	→ /ma-teka-ʔ/	'famous, well known'
'receive'	/n-topu/	+ /ma-…-ʔ/	→ /ma-topu-ʔ/	'received'
'count'	/n-soʔi/	+ /ma-…-k/	→ /ma-soʔi-k/	'counted'

7. Many words nominalised with /ma-…-ʔ/ have semantics similar to adjectives in other languages. However, there is no morphosyntactic basis for distinguishing separate words classes of nouns and adjectives (Edwards 2020: 259).

Table 14. *(continued)*

Root gloss	Stem	Circumfix	Derivation	Gloss
'fold'	/n-riʔi/	+ /ma-…-k/	→ /ma-riʔi-k/	'folded'
'grate'	/n-fona/	+ /ʔ-…-ʔ/	→ /ʔ-fona-ʔ/	'grater'
'bind'	/n-futu/	+ /ʔ-…-ʔ/	→ /ʔ-futu-ʔ/	'cloth band'
'sit'	/n-toko/	+ /ʔ-…-ʔ/	→ /ʔ-toko-ʔ/	'chair'
'sweep'	/n-sapu/	+ /ʔ-…-ʔ/	→ /ʔ-sapu-ʔ/	'broom'

Table 15. Nominalising /ma-…-ʔ/ and /ʔ-…-ʔ/, CVV# roots

Root gloss	Stem	Circumfix	Derivation	Gloss
'writes'	/n-tui/	+ /ma-…-ʔ/	→ /ma-tu‹ʔ›i/	'written'
'is aware'	/na-keo/	+ /ma-…-ʔ/	→ /ma-ke‹ʔ›o/	'aware, cognisant'
'believes'	/n-pirsai/	+ /ma-…-ʔ/	→ /ma-pirsa‹ʔ›i/	'believing'
'writes'	/n-tui/	+ /ʔ-…-ʔ/	→ /ʔ-tu‹ʔ›i/	'pen'
'covers'	/n-neo/	+ /ʔ-…-ʔ/	→ /ʔ-ne‹ʔ›o/	'umbrella, covering'
'pounds'	/na-pau/	+ /ʔ-…-ʔ/	→ /ʔ-pa‹ʔ›u/	'mortar and pestle'
'wife'	/fee/	+ /ma-…-ʔ/	→ /ma-fe‹ʔ›e/	'having a wife'
'leaf'	/noo-f/	+ /ma-…-ʔ/	→ /ma-no‹ʔ›o/	'leafy'
'sews'	/n-soo/	+ /ma-…-ʔ/	→ /ma-so‹ʔ›o/	'sewn, stitched'
'base'	/uu-f/	+ /ma-…-ʔ/	→ /ma-ʔu‹ʔ›u/	'based'
'sings'	/n-sii/	+ /ʔ-…-ʔ/	→ /ʔ-si‹ʔ›i/	'song'

5.2.2 *No evidence of being inalterable (criterion (1b)): Plural enclitic allomorphy and consonant insertion*

Another morphological phenomenon which sheds light on the status of putative double vowels is the behaviour of the different allomorphs of the plural enclitic. This enclitic has different allomorphs depending on the shape of the stem to which it attaches and the word class of the stem. Data in this section is from Kotos Amarasi and we only discuss allomorphs which impact on the analysis of putative double vowels.[8]

On consonant-final stems (both nouns and verbs), the plural enclitic usually has the form /=ein/ ~ /=eni/, and the stem has final CV → VC metathesis. (See Section 5.2.3 for more discussion of metathesis.) After stems which end in CV, the

8. See Edwards (2020: 234–244) for full discussion of the allomorphy of the plural enclitic in Kotos Amarasi.

plural enclitic usually takes the form /=n/. Examples are given in Table 17. Similarly, after verbs which end in a vowel sequence (including a putative double vowel) the plural enclitic also has the form /=n/. Examples of each pattern are given in Table 16–Table 18.

Table 16. Plural enclitic on CVC-final stems

Stem		Plural	Gloss
/anah/	→	/aanh=ein/ 'children'	'children'
/kaes mutiʔ/	→	/kaes muitʔ=ein/	'Europeans'
/enoʔ/	→	/eonʔ=ein/	'doors'
/tuaf/	→	/tuaf=ein/	'people'
/kuan/	→	/kuan=ein/	'villages'
/n-fesat/	→	/n-feest=ein/	'(they) throw a party'
/na-barab/	→	/na-baarb=ein/	'(they) prepare'
/n-ʔonen/	→	/n-ʔoenn=ein/	'(they) pray'
/na-tuin/	→	/na-tuin=ein/	'(they) follow'

Table 17. Plural enclitic on CV-final stems

Stem		Plural	Gloss
/kase/	→	/kase=n/	'foreigners'
/hutu/	→	/hutu=n/	'head-lice'
/kbiti/	→	/kbiti=n/	'scorpions'
/koro/	→	/koro=n/	'birds'
/tuni/	→	/tuni=n/	'eels'
/n-moʔe/	→	/n-moʔe=n/	'(they) do/make'
/na-tona/	→	/na-tona=n/	'(they) tell'
/n-eki/	→	/n-eki=n/	'(they) bring'
/na-hana/	→	/na-hana=n/	'(they) cook'

However, after nominals which end in a vowel sequence, as well as nominals which end in a putative double vowel the plural enclitic has the form /=ngʷein/ [ŋgwɪn] in Kotos Amarasi. Examples are given in Table 19. Other varieties of Meto have different allomorphs after nominals with a final vowel sequence (e.g. Amanuban *=nuu*, Amfo'an *=nuug*), but their distribution is identical to Kotos Amarasi (Culhane & Edwards 2025).

Table 18. Plural enclitic on VV-final verbal stems

Stem		Plural	Gloss
/n-nao/	→	/n-nao=n/	'(they) go'
/n-sii/	→	/n-sii=n/	'(they) sing'
/na-mnau/	→	/na-mnau=n/	'(they) remember'
/n-murai/	→	/n-murai=n/	'(they) start'
/n-tui/	→	/n-tui=n/	'(they) write'
/n-kae/	→	/n-kae=n/	'(they) cry'
/na-niu/	→	/na-niu=n/	'(they) bathe'
/n-poi/	→	/n-poi=n/	'(they) exit/go out'

Table 19. Plural enclitic on VV-final nominal stems

Stem		Plural	Gloss
/bidʒae/	→	/bidʒae=ŋgʷein/	'cows'
/oe/	→	/oe=ŋgʷein/	'kinds of water'
/hau/	→	/hau=ŋgʷein/	'trees'
/hai/	→	/hai=ŋgʷein/	'ours (excl.)/our things'
/bifee/	→	/bifee=ŋgʷein/	'women'
/too/	→	/too=ŋgʷein/	'citizens'
/hoo/	→	/hoo=ŋgʷein/	'yours (sg.)/your things'
/hii/	→	/hii=ŋgʷein/	'yours (pl.)/your things'

Recall that the Meto foot is a weight-insensitive trochee (Section 5.1) and that vowel sequences do not form heavy syllables. Thus, the data in (5) cannot be explained as being due to the foot structure.

The most straightforward analysis is to propose that the allomorphy of the plural enclitic is dependent on the shape and word class of the stem, and that verbs with final vowel sequence take the allomorph /=n/, while nominals with a final vowel sequence take the allomorph /=ŋgʷein/. Another morphological process in Meto which provides evidence that putative double vowels behave identically to other vowel sequences comes from word-final consonant insertion. In several varieties of Meto, vowel-final nouns have a consonant inserted in certain syntactic phrase positions. Such consonant insertion is known to occur in Amfo'an, Timaus, Baikeno, Fatule'u, Kopas, and some varieties of Molo. In some of these (Amfo'an and Timaus), consonants are inserted after almost all vowel-final nouns. In other varieties of Meto, such as Baikeno, it only occurs after nouns

with a final vowel sequence.[9] In Baikeno /ʤ/ is inserted after /Vi/, /l/ is inserted after /Ve/, and /b/ is inserted after /Vo/ and /Vu/. Examples of consonant insertion in Baikeno are given in Table 20. Table 20 includes Proto-Rote-Meto (PRM) etyma (Edwards 2021), Proto-Malayo-Polynesian (PMP) etyma (Blust & Trussel 2020), where these have been reconstructed, as well as Kotos Amarasi cognates for comparison. The Baikeno nouns are in the citation form, one phrase position in which consonant insertion occurs. As stated previously, the Meto foot is a weight insensitive trochee and stress falls on the penultimate syllable for such words, as it does for all other words, e.g. /noel/ → [ˈnɔ.ɛl] 'river'.

Table 20. Baikeno consonant insertion

PMP	PRM	Kotos	Baikeno	Meto gloss
*taqi	*tai	/tei/	/teiʤ/	'faeces'
*waRi	*fai	/fai/	/faiʤ/	'night'
*hapuy	*ahi	/ai/	/aiʤ/	'fire'
*bukij	*fui	/fui/	/fuiʤ/	'wild'
*wahiR	*oe	/oe/	/oel/	'water'
	*lee	/noe/	/noel/	'river'
		/bukae/	/bukael/	'provisions'
*kahiw	*kaiu	/hau/	/haub/	'wood, tree'
*qihu	*iu	/iik iu/	/iub/	'shark, whale'
*qapuR	*aho	/ao/	/aob/	'mineral lime'
	*meo	/meo/	/meob/	'cat'
*kəmiq	*ka-mii	/kmii/	/miiʤ/	'urine'
*di-wiRi	*ka-ɗii	/ʔrii/	/a-ʔliiʤ/	'left'
*bahi	*fee	/bifee/	/bifeel/	'woman'
*bahi	*fee	/fee/	/feel/	'wife'
*tau	*tou	/too/	/toob/	'population'

As can be seen from the comparative data, the final consonants in Baikeno are historic insertions. In Baikeno, as in other varieties of Meto with such consonant insertion, consonant final forms alternate with vowel-final forms. Vowel

9. Baikeno words ending in CV# do not undergo such consonant insertion. Examples include (among many others): *punti > *hundi > /uki/ 'banana', *pajay > *hade > /ane/ 'field rice', *qalejaw > *ledo > /neno/ 'day, sky', and *batu > *batu > /fatu/ 'rock, stone', all of which are vowel-final in the citation form.

final forms occur before an attributive nominal modifier, while consonant final forms occur elsewhere, including in citation form. Examples of Baikeno nouns in Table 20 before attributive modifiers include /hau uu-f/ 'tree trunk', /ai maputuʔ/ 'hot fire', and /bifee banuʔ/ 'widow'.

There are various ways in which this consonant insertion could be analysed. For the purposes of this paper, it is sufficient to note that words with a putative final double vowel undergo consonant insertion in the same way as words with a final vowel sequence. Because consonant insertion and plural enclitic allomorphy treat putative double vowels the same way as other vowel sequences, there is no evidence of them being inalterable (criterion (1b)).

5.2.3 *Double vowels arise from a morphological process (criterion (1d)): Metathesis*

Another feature of Meto which provides support for the double vowels analysis is metathesis. Meto has a productive process of final CV → VC metathesis which is morphological in some cases and phonologically conditioned in others, see Edwards (2020) for details. Following a convention established in that work, we refer to unmetathesized forms and paradigmatically equivalent forms as *U-forms*, while metathesized forms and those forms paradigmatically equivalent to them are referred to as *M-forms*.

Disyllabic roots with the shape (C)$V_{\alpha}CV_{\beta}$ provide the clearest examples. Such roots undergo final CV → VC metathesis, resulting in a derived form with a vowel sequence. Examples of metathesis from Kotos Amarasi are given in Table 21 to illustrate. As with all other words of the language, the resulting metathesized form is a weight-insensitive trochee with stress on the penultimate syllable, e.g. /neno/ ['nɛnɔ] → /neon/ ['nɛ.ɔn] 'day; sky'.

Table 21. Kotos Amarasi metathesis (Edwards 2020: 162)

U-form	M-form	Gloss	U-form	M-form	Gloss
/be**si**/	/be**is**/	'knife'	/ne**no**/	/ne**on**/	'day; sky'
/fa**fi**/	/fa**if**/	'pig'	/kna**fo**/	/kna**of**/	'mouse'
/o**ni**/	/o**in**/	'bee'	/hi**tu**/	/hi**ut**/	'seven'
/u**ki**/	/u**ik**/	'banana'	/te**nu**/	/te**un**/	'three'
/ba**re**/	/ba**er**/	'place'	/fa**tu**/	/fa**ut**/	'stone'
/no**pe**/	/no**ep**/	'cloud'	/no**pu**/	/no**up**/	'hole, grave'

For roots with identical penultimate and final vowels, as well as roots with /a/ in the final syllable, the M-form has a double vowel in most varieties of Meto.[10] Examples from Kotos Amarasi are given in Table 22.

Table 22. Kotos Amarasi metathesized forms with putative double vowel

U-form	M-form		Gloss
/fi**ni**/	/fi**in**/	[fiːn]	'seed'
/na-ski**ki**/	/na-ski**ik**/	[naˈskiːk]	'brushes teeth'
/re**ne**/	/re**en**/	[rɛːn]	'field'
/ume**ke**/	/ume**ek**/	[ʊˈmɛːk]	'k.o. brown snake'
/pa**ha**/	/pa**ah**/	[paːh]	'country'
/na-ta**na**/	/na-ta**an**/	[naˈtaːn]	's/he asks'
/ko**ro**/	/ko**or**/	[kɔːr]	'bird'
/na-bso**ʔo**/	/na-bso**oʔ**/	[nabˈsɔːʔ]	's/he dances'
/hu**tu**/	/hu**ut**/	[huːt̪]	'head louse'
/tung^{w}**uru**/	/tung^{w}**uur**/	[t̪ʊŋˈguːr]	'teacher'

Under a double vowel analysis, the derived form for roots with identical penultimate and final vowels in Table 22 can be straightforwardly analysed as undergoing metathesis in the same way as words with differing vowels; the final vowel and consonant have exchanged positions resulting in a sequence of two identical vowels. Whether there are phonetic differences between double vowels which have arisen through metathesis and double vowels which occur in roots is an open question for future research.

5.3 Analytical parsimony (criterion (1e))

A double vowel analysis of vowel length in Meto can also be made on the basis of analytical parsimony. Recall from Section 3 that Meto does not have segmental diphthongs, and that all diphthongs are analysed as underlying sequences of two monophthongs. Meto attests every sequence of two vowels with the exception of a high vowel and mid vowel of the same backness (*ie, *uo), as shown in Table 23a.[11]

10. Kusa-Manea preserves final /a/ after metathesis for a-final forms such as /nima/ → /niam/ [ˈni.am] 'five'. It has double vowels in metathesized forms with identical penultimate and final vowels e.g. /na-hini/ → /na-hiin/ [naˈhiːn] 'knows.

11. Some varieties of Meto prohibit all sequences of a high vowel and mid vowel: *ie, *io, *uo, and *ue.

Double vowels fit within the system of permitted vowel sequences — i.e. no additional stipulations are required to account for their occurrence. This is as shown in Table 23b. Additionally, under a double vowel analysis, the distribution of long vowels is the same as for other sequences of two vowels and no additional phonotactic restrictions required to account for their distribution (see Section 5.1). On the other hand, if we were to analyse Meto as having segmental long vowels, this would mean doubling the size of the vowel inventory, and proposing a more complex account of their phonotactic distribution (as discussed in Section 5.1).

Table 23. Sequences of two monophthongs attested in Meto (Amfo'an)

a. Without double vowels

V1 ↓	**i**	**e**	**a**	**o**	**u**	**V2 →**
i			ia	io	iu	
e	ei		ea	eo	eu	
a	ai	ae		ao	au	
o	oi	oe	oa		ou	
u	ui	ue	ua			

b. With double vowels

V1 ↓	**i**	**e**	**a**	**o**	**u**	**V2 →**
i	ii		ia	io	iu	
e	ei	ee	ea	eo	eu	
a	ai	ae	aa	ao	au	
o	oi	oe	oa	oo	ou	
u	ui	ue	ua		uu	

5.4 Summary

In this section, we have surveyed phenomena which bear upon the phonological analysis of vowel length in Meto. We have shown that long vowels in Meto demonstrate no evidence of integrity (criterion [1a]), but undergo glottal stop infixation like other sequences of vowels (Section 5.2.1). Long vowels in Meto also demonstrate no evidence of being inalterable (criterion [1b]), but demonstrate the same morphophonological behaviour as other sequences of two vowels. Their phonotactic distribution is also the same as sequences of two vowels (criterion

[1c]) and they arise from morphological processes (criterion [1d]). Further, it is in the interests of analytical parsimony to analyse Meto as having double vowels (criterion [1e]). Taken together, these findings provide strong support for analysing Meto as having double vowels.

6. Discussion and conclusion

The results of the current study reveal that there is acoustic evidence for analysing Meto as having a vowel length distinction. Putative double vowels are approximately 1.4 times as long as single vowels, and there are some durational differences between consonants following double and single vowels. There is also some evidence of vowel quality differences between single and double vowels.

Employing a range of criteria used to distinguish between long single segments from sequences of identical segments, we find that Meto is best analysed as having double vowels rather than segmental long vowels. Evidence for this analysis comes from their phonotactic distribution, morphophonological behaviour and analytical parsimony: double vowels have the same phonotactic distribution and (morpho)phonological behaviour as unambiguous sequences of two vowels.

A double vowel analysis has several implications for phonological theory. Most significantly, it is a violation of the Obligatory contour principle (OCP). The OCP is a constraint which forbids representations in which identical elements are adjacent (Leben 1973; Goldsmith 1976; McCarthy 1986). Apparent counterexamples abound (Odden 1986; Myers 1997) and it is no longer assumed that the OCP is a universal principle, but is a cross-linguistic tendency and, therefore, a violable constraint (Guy & Boberg 1997; Myers 1997).

These findings support previous descriptions of Meto as having a phonological distinction between single and double vowels and provide us with an understanding of the acoustic cues associated with this distinction. They also highlight that there is not a transparent relationship between the phonetic cues of vowel length and its phonological patterning: while double vowels in Meto behave as sequences of two monophthongs, double vowels are only ~40% longer than single vowels; and the distinction is also associated — to some extent — with differences in vowel quality and consonant duration.

More broadly, this study highlights how field data can be used in analysis. Here, we have conducted an acoustic study of vowel length using spontaneous data collected in a non-ideal setting; normal rooms in an Indonesian village. It was also recorded using fairly basic recording equipment: the in-built microphone of a Zoom H1 recorder. While spontaneous data has weaknesses, and is not always suitable (see Xu 2010; Wagner et al. 2015), using such data allows us to con-

duct research on languages which are otherwise not represented in the phonetics literature.

The analytical choices outlined in this study also have flow-on effects. Firstly, it affects the transcription of Meto in linguistic work whereby double vowels are typically transcribed with two symbols (e.g. Edwards 2020; Culhane 2018; Tan 2022a, 2022b, 2023; Lemon 2024). Secondly, they have informed the development of the practical orthography of varieties of Meto. In local materials double vowels are now written with two letters, e.g. <*bifee*> /bifee/ [bɪˈfɛː] 'woman', <*haa*> /haa/ [haː] 'four'. This goes against the initial instinct of some "naïve" native speakers, and the practice of earlier linguists (e.g. Middelkoop 1939, 1950; Steinhauer 1993, 1996a, 1996b), whereby double vowels are often transcribed with a single letter.[12] Our discussions with teachers and native speakers indicate that once introduced to the concept of writing double vowels, native speakers of Meto consistently find materials written with two letters easier to read and understand than those without.

While we are not the first to analyse a language as having double vowels (e.g. Jakobi 1990; Zwarts 2004; Treis 2008; Heston 2015; McCracken 2019; Forker 2020), this paper highlights the need to reflect on the analytical decisions made (or assumed) during the documentation process, to be transparent about these decisions, to revise them as needed. We thus underscore the work of others who have pointed out that atheoretical language documentation is a fallacy (Himmelmann 1998, 2006; Gil 2001; Dryer 2006); as any transcription implies some level of phonological and morphological analysis.

Himmelmann (1998, 2006) among others rightly states that the goal of language documentation is to compile comprehensive, representative and lasting records of natural language. In the case of Meto, creating accurate and comprehensive records of the language requires that the vowel length distinction is taken into account, phonological analysis undertaken, and analytical decisions justified.[13] Indeed, a lack of in-depth analysis and transparency about the decisions made can mean that data is misinterpreted or its future usability is restricted (Klamer 2009: 251).

12. For native speakers, this practice is based on the model of Indonesian orthography which does not distinguish between vowels of differing lengths. The reason for this practice in older linguistic work is not explained.

13. Not considered in this paper is the morphological analysis also needed. This too has impacts on orthography, most obviously in regards to the placement of word breaks.

Acknowledgments

We would like to thank two anonymous reviewers whose comments on an earlier version of this paper have improved the presentation and discussion. The analyses in this paper builds upon that developed by others including, but not limited to, Heronimus Bani and Charles E. Grimes. We also wish to express our appreciation to Nikolaus Himmelmann, both for this mentorship and commitment to the documentation of minority languages of Indonesia.

Abbreviations

Glossing follows the Leipzig Glossing Rules.

References

Abramson, Arthur S., and Nianqi Reo. 1990. "Distinctive Vowel Length: Duration vs. Spectrum in Thai". *Journal of Phonetics* 18: 79–92.

Barnes, Jonathan Allen. 2002. "Positional neutralization: A phonologization approach to typological patterns." PhD thesis, University of California, Berkeley. ProQuest (288390119).

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models using lme4." *Journal of Statistical Software* 67: 1–48.

Bauer, Winifred. 1993. *Maori*. London and New York: Routledge.

Beckman, Jill N. 1998. "Positional Faithfulness". PhD thesis, University of Massachusetts, Amherst. ProQuest (304424695).

Blust, Robert, and Stephen Trussel. 2020. "The Austronesian Comparative Dictionary." http://www.trussel2.com/ACD/

Boersma, Paul, and David Weenink. 2023. "Praat: Doing Phonetics by Computer." https://praat.org/

Brand, James, Jen Hay, Lynn Clark, Kevin Watson, and Márton Sóskuthy. 2021. "Systematic Co-Variation of Monophthongs across Speakers of New Zealand English." *Journal of Phonetics* 88: 101096.

Chen, Chun-Mei. 2006. "A Comparative Study on Formosan Phonology: Paiwan and Budai Rukai." PhD thesis, The University of Texas at Austin. ProQuest (304981925).

Chen, Matthew. 1970. "Vowel Length Variation as a Function of the Voicing of the Consonant Environment." *Phonetica* 22: 129–159.

Culhane, Kirsten. 2017. "Amfo'an Recordings (KC1)." *Paradisec*. http://catalog.paradisec.org.au/collections/KC1

Culhane, Kirsten. 2018. "Consonant Insertions: A Synchronic and Diachronic Account of Amfo'an." BA Honours thesis, The Australian National University.

Culhane, Kirsten. 2022. "The Phonology and Typological Position of Waima'a Consonants." *Oceanic Linguistics* 61: 528–559.

Culhane, Kirsten. 2024. "The Prosodic Foot: A Typological Study of Greater Timor Languages." PhD thesis, University of Freiburg.

Culhane, Kirsten, and Owen Edwards. 2025. "Consonant Epenthesis in Meto: Typologically Rare but Diachronically Explicable." In *Rarities in Phonetics and Phonology*, ed. by Natalia Kuznetsova, Cormac Anderson, and Shelece Easterday 167–219. Berlin: Language Science Press.

Dmitrieva, Olga. 2012. "Geminate Typology and the Perception of Consonant Duration." PhD thesis, Stanford University.

Dryer, Matthew S. 2006. "Descriptive Theories, Explanatory Theories, and Basic Linguistic Theory." In *Catching Language: The Standing Challenge of Grammar Writing*, ed. by Alan Dench, Felix K. Ameka, and Nicholas Evans, 207–234. Berlin: De Gruyter Mouton.

Edwards, Owen. 2017. "Epenthetic and Contrastive Glottal Stops in Amarasi." *Oceanic Linguistics* 56: 415–434.

Edwards, Owen. 2018. "The Morphology and Phonology of Metathesis in Amarasi." *Morphology* 28: 25–69.

Edwards, Owen. 2020. *Metathesis and Unmetathesis in Amarasi*. Berlin: Language Science Press.

Edwards, Owen. 2021. *Rote-Meto Comparative Dictionary*. Canberra: ANU Press.

Esposito, Anna, and Maria Gabriella Di Benedetto. 1999. "Acoustical and perceptual study of gemination in Italian stops." *The Journal of the Acoustical Society of America* 106: 2051–2062.

Forker, Diana. 2020. *A Grammar of Sanzhi Dargwa*. Berlin: Language Science Press.

Gil, David. 2001. "Escaping Eurocentrism: Fieldwork as a Process of Unlearning." In *Linguistic Fieldwork*, ed. by Paul Newman, and Martha Ratliff, 102–132. Cambridge: Cambridge University Press.

Goldsmith, John. 1976. "Autosegmental Phonology." PhD thesis, MIT.

Guy, Gregory R., and Charles Boberg. 1997. "Inherent Variability and the Obligatory Contour Principle." *Language Variation and Change* 9: 149–164.

Harlow, Ray. 2007. *Maori: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Harvey, Mark, Nay San, Margaret Carew, Sydney Strangways, Jane Simpson, and Clara Stockigt. 2019. "Pre-Stopping in Arabana." *Australian Journal of Linguistics* 39: 419–462.

Hayes, Bruce. 1986a. "Assimilation as Spreading in Toba Batak." *Linguistic Inquiry* 17: 467–499. https://www.jstor.org/stable/4178500

Hayes, Bruce. 1986b. "Inalterability in CV Phonology." *Language* 62(2): 321–351.

Hellwig, Birgit. 2019. *A Grammar of Qaqet*. Berlin and Boston: De Gruyter Mouton.

Heston, Tyler. 2015. "The Segmental and Suprasegmental Phonology of Fataluku." PhD thesis, University of Hawai'i at Mānoa. http://hdl.handle.net/10125/51090

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36: 161–195.

Himmelmann, Nikolaus P. 2006. "Language Documentation: What is it and what is it good for?" In *Essentials of language documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin and New York: De Gruyter Mouton.

Hui, C. T. Justine, and Takayuki Arai. 2019. "Elderly Listeners' Identification of Japanese Long Vowel Pair 'obasan' and 'obaasan' using Pitch and Duration." *Acoustical Science and Technology* 40: 105–115.

Idemaru, Kaori, and Susan G. Guion. 2008. "Acoustic covariants of Length Contrast in Japanese Stops." *Journal of the International Phonetic Association* 38: 167–186.

Jakobi, Angelika. 1990. *A Fur Grammar: Phonology, Morphophonology, and Morphology*. Hamburg: Helmut Buske. https://hdl.handle.net/11858/00-001M-0000-0012-9842-5

Jepson, Kathleen. 2019. "Prosody, Prominence and Segments in Djambarrpuyŋu." PhD thesis, The University of Melbourne.

Jepson, Kathleen. 2025. "Contrastive Vowel Length and Segment Duration in Production and Perception in Djambarrpuyŋu." *Laboratory Phonology* 16: 1–49.

Kaland, Constantijn. 2021. "Praat Scripts." https://osf.io/cr4vn/

Keating, Patricia Ann. 1985. "Universal Phonetics and the Organization of Grammars." In *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, ed. by Victoria Fromkin, 115–132. Orlando: Academic Press.

Kehrein, Wolfgang. 2002. *Phonological Representation and Phonetic Phasing: Affricates and Laryngeals*. Tübingen: Max Niemeyer.

Kenstowicz, Michael J., and Charles Pyle. 1973. "On the Phonological Integrity of Geminate Clusters." In *Issues in Phonological Theory*, ed. by Michael J. Kenstowicz, and Charles W. Kisseberth, 27–43. Berlin and Boston: De Gruyter Mouton.

Klamer, Marian. 2009. "The use of language data in comparative research: A note on Blust (2008) and Onvlee (1984)." *Oceanic Linguistics* 48: 250–263.

Kluender, Keith R., Randy L. Diehl, and Beverly A. Wright. 1988. "Vowel-length differences before voiced and voiceless Consonants: an auditory explanation." *Journal of Phonetics* 16: 153–169.

Lahiri, Aditi, and Jorge Hankamer. 1988. "The timing of geminate consonants." *Journal of Phonetics* 16: 327–338.

Leben, William R. 1973. "Suprasegmental Phonology." PhD thesis, MIT. https://hdl.handle.net/11858/00-001M-0000-0025-7021-0

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, Ilse. 1976. "Influence of fundamental frequency pattern on the perception of duration." *Journal of Phonetics* 4: 113–117.

Lehnert-LeHouillier, Heike. 2007. "The influence of dynamic f0 on the perception of vowel duration: Cross-linguistic evidence." *Proceedings of the 16th International Congress of Phonetic Sciences (ICPHS XVI)*. Universität des Saarlandes.

Lemon, Tyler. 2024. "The Morphosyntax of Verbal Agreement in Uab Meto." PhD thesis, University of California, Berkeley. ProQuest (Lemon_berkeley_0028E_23660).

Lenth, Russell V. 2024. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.6, https://rvlenth.github.io/emmeans/

Maddieson, Ian. 1985. "Phonetic cues to syllabification." In *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, ed. by Victoria Fromkin, 203–221. Orlando: Academic Press.

McCarthy, John. 1986. "OCP effects: Gemination and Antigemination." *Linguistic Inquiry* 17: 207–263.

McCracken, Chelsea. 2019. *A Grammar of Belep*. Berlin and Boston: De Gruyter Mouton.

Middelkoop, Pieter. 1939. "*Amarassische-Timoreesche Teksten*." Bandung: Verhandelingen van het Bataviaasch Genootschap voor Kunsten en Wetenschappen 74 (2).

Middelkoop, Pieter. 1950. "Proeve van een Timorese grammatica." *Bijdragen tot de Taal-, Land- en Volkenkunde* 106: 375–517.

Mooney, Kate. 2022. "Phonology Cannot Transpose: Evidence from Meto." *Phonology* 39: 293–343.

Myers, Scott. 1997. "OCP Effects in Optimality Theory." *Natural Language and Linguistic Theory* 15: 847–892.

Myers, Scott, and Benjamin B. Hansen. 2007. "The Origin of Vowel Length Neutralization in Final Position: Evidence from Finnish Speakers." *Natural Language and Linguistic Theory* 25: 157–193.

Odden, David. 1986. "On the Role of the Obligatory Contour Principle in Phonological Theory." *Language* 62: 353–383.

Odden, David. 2011. "The Representation of Vowel Length." In *The Blackwell Companion to Phonology*, ed. by Marc van Oostendorp, Colin Ewen, Elizabeth Hume, and Keren Rice. Chichester: Wiley-Blackwell.

Paschen, Ludger, Susanne Fuchs, and Frank Seifart. 2022. "Final lengthening and vowel length in 25 languages." *Journal of Phonetics* 94: 101179.

von Prince, Kilu. 2015. *A grammar of Daakaka*. Berlin: De Gruyter Mouton.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Remijsen, Bert, and Leoma Gilley. 2008. "Why are Three-Level Vowel Length Systems Rare? Insights from Dinka (Luanyjang Dialect)." *Journal of Phonetics* 36: 318–344.

Ridouane, Rachid. 2007. "Gemination in Tashlhiyt Berber: An Acoustic and Articulatory Study." *Journal of the International Phonetic Association* 37: 119–142.

van Santen, Jan P. H. 1992. "Contextual Effects on Vowel Duration." *Speech Communication* 11: 513–546.

Steinhauer, Hein. 1993. "Notes on Verbs in Dawanese." In *Topics in Descriptive Austronesian Linguistics*, ed. by Ger Reesink, 130–158. Leiden: Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië.

Steinhauer, Hein. 1996a. "Morphemic Metathesis in Dawanese (Timor)." In *Papers in Austronesian Linguistics 3*, ed. by Hein Steinhauer, 217–232. Canberra: Pacific Linguistics. https://hdl.handle.net/21.11116/0000-000D-2452-E

Steinhauer, Hein. 1996b. "Synchronic Metathesis and Apocope in Three Austronesian Languages of the Timor Area." In *Pan-Asiatic Linguistics; Proceedings on the Fourth International Symposium on Language and Linguistics, 8-10 January*, ed. by Suwilai Premsrirat, 471–492. Thailand: Institute of Language and Culture for Rural Development, Mahidol University. https://hdl.handle.net/2066/67875

Summers, W. Van. 1987. "Effects of Stress and Final-Consonant Voicing on Vowel Production: Articulatory and Acoustic Analyses." *The Journal of the Acoustical Society of America* 82: 847–863.

Tabain, Marija, and Birgit Hellwig. 2023. "Qaqet." *Journal of the International Phonetic Association* 53: 1073–1094.

Tan, Tamisha. 2022a. "Amarasi Texts and Meto Elicitation Recordings (TLT1)." *Paradisec*: https://catalog.paradisec.org.au/collections/TLT1

Tan, Tamisha. 2022b. "Two Cases of Doubled Pronouns in Amarasi." *Journal of East Asian Linguistics* 31: 551–590.

Tan, Tamisha. 2023. "Voice and Valency in Amarasi: Topics in Synchronic and Diachronic Morphosyntax." PhD thesis, Harvard University.

Treis, Yvonne. 2008. *A Grammar of Kambaata, 1: Phonology, Nominal Morphology and Non-Verbal Predication*. Köln: Rüdiger Köppe.

Trubetzkoy, Nikolai S. 1939. *Grundzüge der Phonologie*. Prague: Travaux du Cercle Linguistique de Prague.

Wagner, Petra, Jürgen Trouvain, and Frank Zimmerer. 2015. "In defense of stylistic diversity in speech research." *Journal of Phonetics* 48: 1–12.

Wilkinson, Melanie P. 1991. "Djambarrpuyŋu: A Yolŋu Variety of Northern Australia." PhD. thesis, University of Sydney. https://hdl.handle.net/21.11116/0000-000D-21C7-D

Wilson Black, Joshua, and James Brand. 2022. *nzilbb.vowels: Functions for Vowel Covariation Studies*. https://github.com/JoshuaWilsonBlack/nzilbb_vowels/tree/main

Xu, Yi. 2010. "In defense of lab speech." *Journal of Phonetics* 38: 329–336.

Zwarts, Joost. 2004. *The Phonology of Endo: A Southern Nilotic Language of Kenya*. Lincom.

CHAPTER 18

Prosodic effects on glottal stop realization and voice quality variation in a corpus study of Tahitian

Janet Fletcher & Adele Gregory
University of Melbourne

Languages with "simple" phonological inventories are often neglected in traditional descriptive phonetic research in favour of those with more varied segmental and prosodic features. Tahitian (Eastern Polynesian) is a prime example with no experimental phonetic investigation of the language in spite of a relatively long history of language documentation. This chapter explores the interaction between prosody (stress, phrasing), glottal stop variation and vowel glottalization. Phonemic glottal stops in many languages of the world have a range of variants, from fully occluded voiceless glottal stops to creaky voiced vowels. Tahitian is no exception but there is a relatively complex relationship between glottal variation and prosodic function (i.e. prominence and phrasing), suggesting that languages with "simple" phonological systems illustrate phonetic complexity.

Keywords: Tahitian, glottal stop, glottalization, stress

1. Introduction

In the last twenty years or more there has been significant uplift in the development of tools and methodologies to facilitate large scale linguistic analyses of fieldwork corpora to advance language documentation. The work of Himmelmann and colleagues (e.g. Himmelmann 1998, 2006; Gippert et al. 2006) has been squarely in the vanguard of this development. A strong focus of Himmelmann's has been the importance of recording multi-genre corpora to enable refinement of core linguistic description of under-resourced languages (e.g. see Himmelmann 1998: 162). In a similar vein, there has been significant expansion in large-scale quantitative cross-linguistic phonetic and phonological analyses of under-resourced or endangered languages (see for example the overview in Whalen & McDonough 2015; Gordon 2016; Whalen et al. 2022). Indeed, it is probably in

https://doi.org/10.1075/slcs.240.18fle
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

the area of prosodic description and prosodic typology where the agendas of language documentation and cross-linguistic phonetic and phonological research most coincide and this has been a further important focus of Himmelmann's among others (e.g. Himmelmann & Ladd 2008; Jun & Fletcher 2014; Di Canio & Bennett 2020; Gussenhoven & Chen 2020 and chapters therein). An important additional methodological concern in crosslinguistic phonetic and phonological research has been the need to elicit a range of experimentally controlled materials including controlled sentence frame elicitation tasks, readings of a standard passage e.g. *The Northwind and the Sun* fable (e.g. see the discussion in Ladefoged & Maddieson 1996; Maddieson 2001).

The development and wide-spread availability of forced alignment tools such as WebMaus (Kisler et al. 2017) and the Montral Forced Aligner MFA (McAuliffe et al. 2017) has also meant that there are increasingly large scale phonetic analyses of fieldwork datasets that are part of research initiatives including the DoReCo (Language Documentation Reference Corpora) project (e.g. Blum et al. 2024; Paschen et al. 2022), that have allowed investigation of key phonetic typological patterns across large naturalistic datasets. Other recent corpus analyses have inspired re-evaluation of phonetic typology (e.g. see Chodroff 2025 for a good summary). Moreover, the use of intonation contour clustering methodologies by Kaland and colleagues will hopefully see further uplift in studies of intonational variation across fieldwork languages (Kaland 2023). This ongoing refinement of phonetic universals and phonetic typology goes beyond the development of new analytical tools, however, and has been in train for some years (e.g. Gordon 2016).

The focus of this chapter is an exploration of the interaction between segmental realization and prosodic structure in Tahitian *Reo Tahiti*, an Eastern Polynesian language which is the dominant indigenous language of the French Polynesian archipelago. It is estimated that around 70% of the population of French Polynesia speak, read, or write Tahitian or one of the other six indigenous languages of the archipelago. Tahitian is the most widespread in terms of daily usage although undergoing a process of reduced transmission from older to younger generations, particularly in Tahiti itself (Paia & Vernaudon 2002; Walworth 2020). There has been severe language loss and at least four of the six other languages are deemed endangered due to a process of "Tahitianization" as well as cultural intrusion of French since the early 1800s (Walworth 2020; Charpentier & François 2015). Tahitian has a reasonably long history of language documentation and description spanning more than 200 years (e.g. Davies 1851; Jaussen 1886; Tryon 1970; Coppenrath & Prévost 1975; Lemaître 1973; Lazard & Pelzer 1991; Vernaudon 2011; Fare Vāna'a 1999, 2009). Moreover, general exposure to the sound systems of the languages of French Polynesia has been promoted via the development of tools such as the Speaking Atlas of Indigenous Languages of France and its Overseas Territories

(Boula de Mareüil et al. 2019). There have also been initiatives to boost digital archives for Oceanic languages, focussing on the languages of French Polynesia through a specialised PARADISEC initiative (e.g. Vernaudon et al. 2021) and further research initiatives to investigate and document indigenous knowledge and language in Oceania, with a particular focus on French Polynesia, Vanuatu, Wallis and Futuna and New Caledonia (e.g. the HéLiCéO project, Vernaudon et al. 2025–2029).

1.1 Background on Tahitian

In terms of the sound system of Tahitian, there are a number of basic phonological overviews that include an account of stress phenomena as well as segmental contrasts (e.g. Lemaître 1973; Perini 1987; Bodin 2006; Bickmore 1995), but until recently there has been virtually no experimental phonetic study of the sound system, lexical stress or post-lexical prosodic system of Tahitian. To the casual observer, it might appear that languages with smaller or "simpler" phonological inventories like Tahitian are less "interesting" compared to languages with more varied or "exotic" segmental and prosodic features which could explain why Tahitian, with its relatively simple phonemic profile (see Table 1 below) may not have received as much experimental phonetic attention as other languages of the world. While the exploration of diverse languages and complex (as well as more typologically common) speech sounds in seminal works such as Ladefoged and Maddieson (1996) has contributed enormously to our understanding of phonetic diversity and phonetic typology and indeed to phonetic theory in general, it is still nonetheless important to examine languages with less obviously complex phonological contrasts. Another point that needs to be made is that fewer phoneme categories in a language or language variety does not mean there is not a degree of phonetic variation or phonetic complexity within these categories. It has been long accepted in phonetic science that phonetic implementation varies widely across and within languages and instrumental phonetic research has contributed significantly to our understanding of this variation.

Many languages of the Polynesian Triangle which covers a large area of the Pacific Ocean and includes Tahitian as well as Māori, Tongan, Hawaiian, among others, are known for their small consonant inventories with relatively few places of articulation and no voicing contrast in the stop series (e.g. Lynch 1998). Within French Polynesia, consonant inventories range from eight to ten consonants. Tahitian has nine consonants with a stop series consisting of /p,t,ʔ/ as shown in Table 1. This is a relatively typical stop inventory although other languages of French Polynesia, e.g. Marquesan and Pa'umotu, have a voiceless velar phoneme /k/ (Lynch 1998; Mutu & Teìkitutoua 2002; Charpentier & François 2015).

Table 1. Phoneme inventory of Tahitian (after Perini 1987)

Consonants	**Labial**	**Dentialveolar**	**Glottal**
Stops	p	t	ʔ
Nasals	m	n	
Fricatives	f v		h
Liquids		ɾ	

Vowels	**Front**			**Back**
Close	i,iː			u,uː
Mid		e,eː	o,oː	
Open		a,aː		

It is widely agreed in previous phonological descriptions that Tahitian has only open syllables and a quantity-sensitive lexical stress system although descriptions of the latter vary (e.g. Coppenrath & Prévost 1975; Lemaître 1973; Bickmore 1995). According to the most recent account (Bickmore 1995), main stress typically falls on the final syllable if heavy (i.e. contains a long vowel) otherwise it falls on a penultimate syllable if it contains a long vowel or on an antepenultimate syllable that contains a long vowel when the penultimate is a short vowel. Words with light syllables can have variable stress placement varying from penultimate to initial in words longer than two syllables. There has been no major prior analysis of intonational phonology apart from some insights documented by Bickmore (1995), and a preliminary analysis of the intonation system confirms main stressed syllables are potential docking sites for post-lexical intonational pitch accents (Fletcher 2021) suggesting Tahitian is a head prominence language at the phrase level (after Jun 2005).

In this chapter, we focus specifically on the phonetic realization of /ʔ/ given it is an important phoneme in Tahitian, and we also examine glottalization patterns in neighbouring vowels. Phonological words containing glottal stops have been described previously as "[p]honetically eroded" (e.g. Charpentier & François 2015) largely because there has been a process of historic debuccalization with *k and *ŋ now realized as /ʔ/ (Coppenrath & Prévost 1975). We will show by contrast that there is a complex interaction between phonetic implementation of the glottal stop and vowel glottalization variation and prosody (accentual prominence, prosodic position).

Given /ʔ/ is such a pervasive phoneme in Tahitian with high levels of text frequency, its pronunciation has been impressionistically described in earlier phonological overviews but somewhat inconsistently. Bodin (2006) suggests that glottal

stops are always voiceless in Tahitian but suggests there are creaky voiced variants also with unpredictable variation. Perini (1987) suggests there is variation, but that more detailed phonetic analyses are required. Coppenrath and Prévost (1975) also discuss laryngealized (creaky voiced) variants but conclude that variation is somewhat unpredictable. Mutu and Teìkitutoua (2002) describe a tense versus lax variant of /ʔ/ in Marquesan (Ùa pou variety), another Eastern Polynesian language spoken in French Polynesia. They specify the tense allophone as fully stopped and voiceless, whereas the lax allophone is a creaky voiced or "laryngealized" allophone. They also describe neighbouring vowels, particularly vowels following /ʔ/ as being creaky voiced regardless of the variant. A similar observation is made by Lemaître (1973) for Tahitian.

In view of these earlier impressionistic descriptions of Tahitian, a large corpus recorded in Papeete, French Polynesia was analyzed to examine variation in Tahitian /ʔ/ realization. The earlier literature did not specifically mention prosodic context (stress, phrasing) as far as we can tell, so we set out to see whether broad prosodic context is a relevant factor in determining whether there are more voiceless, fully occluded variants in prosodic constituent-initial position, for example or in prosodically prominent syllable onsets. To gain further insights on potential effects of accentual prominence, we also look at vowels flanking /ʔ/ particularly in view of Mutu and Teìkitotoua's (2002) observations for Marquesan and Lemaître's (1973) observations for Tahitian. Before presenting a summary of some quantitative findings, we review some recent research on cross-linguistic glottal stop variation and studies showing links between its implementation and the signalling of prosodic structure. We also review some of the relevant quantitative research on other Eastern Polynesian languages.

1.2 Glottal stop variation and prosodic structure

It is commonly accepted that whilst complete closure of the glottis is the textbook realisation of a glottal stop /ʔ/, there can be variable phonetic implementation within and across the world's languages. Indeed, Ladefoged and Maddieson (1996) suggest that full occlusion is actually very rare particularly in word-medial position. Glottal stop variants range from a period of complete glottal occlusion (voiceless) to a period of creaky voice or even breathy or modal voice along a continuum. Canonical glottal stops are voiceless with zero glottal flow (after Keating et al. 2023). Moreover, a hyperarticulated glottal stop is a complex articulation involving multiple articulatory manoeuvres including incursion of ventricular folds, constriction of epilaryngeal area as well as vocal fold closure and laryngeal raising (see summary in Garellek et al. 2023: 310 after Esling et al. 2019). A fully occluded voiceless glottal stop can also show up acoustically as periods of voicelessness with some

glottal pulsing evident at the margins. Creaky voice also involves complex laryngeal settings with damped and often irregular glottal pulses during the interval of the glottal stop although regular periodicity can also be observed suggesting a high level of gradience (after Keating et al. 2015 and see below).

Prosodic conditioning can also influence glottal stop realisation with more fully occluded, gesturally strengthened variants in Intonational Phrase-initial positions compared to word-medial contexts for example (Kohler 1994; Garellek 2014). This is commensurate with findings for other types of consonants where both lengthening and strengthening can be observed in consonants at the beginning of higher-level prosodic constituents compared to lower-level constituents (Keating et al. 2003). A commonly held assumption is that higher levels of prosodic prominence in general are associated with elevated levels of gestural strengthening with longer consonants, tighter articulatory "seals" in the case of obstruents, and more precisely articulated vowels (see Cho 2016 for an overview). This may not necessarily translate into longer acoustic durations. A recent study of 51 typologically diverse languages from the DoReCo corpus presents cross-linguistic evidence that consonants are lengthened in word-initial positions but not always in utterance-initial position (Blum et al. 2024).

Much has been made of cross-linguistic similarities in patterns of variation in the phonetic implementation of /ʔ/. In a large-scale survey of 131 Illustrations of the International Phonetic Alphabet published in the Journal of the International Phonetic Association, Garellek et al. (2023) showed that acoustic measures of glottal state including percentage of voicing which registers the proportion of voicing during a particular segmental interval, and voicing intensity including strength of excitation measures (SoE) which registers the relative amplitude of pitch pulses, were effective in quantifying differences among glottal stop variants. Creaky voice variants show higher proportions of voicing and stronger levels of voicing intensity compared to partially occluded variants of /ʔ/. Interestingly, this study also highlights the high level of gradience across languages with even utterance initial /ʔ/ showing up to 50% voicing in their corpus. Glottal stops can also show degrees of laryngealization from more "consonantal" creak with intermittent periodicity (irregular F0) and low noise characteristics as well as more periodic "vowel-like" creak or vocal fry (e.g. see Keating et al. 2015 for a detailed overview of different types of creak). Other acoustic measures that are typically used in studies of glottal state variation include different spectral tilt measures (Keating et al. 2023) and noise measures like Harmonics-to-Noise ratios (HNR), which register the noisiness of the signal. Spectral tilt refers to the relationship between the amplitude of lower and higher harmonics in the acoustic spectrum. Lower values are usually indicative of vocal fold tension and constriction and therefore of creaky voice (Garellek 2022).

With respect to quantitative analyses of glottal stop variation in Eastern Polynesian languages, Hawaiian has received the most attention (Parker Jones 2017; Davidson 2021; Davidson & Parker Jones 2023), and to a lesser extent Cook Islands Māori (Nicholas & Coto-Solano 2019). Nicholas and Coto-Solano (2019) found high levels of variation in terms of glottal stop realisation, particularly across the four varieties examined in their study. Speakers of some varieties tend to use more creaky voice variants whereas speakers of other varieties use more canonical glottal stops. As for Hawaiian, Davidson (2021) and Davidson and Parker Jones (2023) examined a corpus of naturalistic speech data from the 1970s and showed (based on visual inspection of spectrograms and waveforms), relatively few instances of /ʔ/ are produced with full glottal closure (7%) with most variants in the corpus realized as creaky voiced variants. In the 2021 study, creaky voiced variants are further subdivided into different sub-variants including a period of creaky voice flanked by two modal vowels: "mid-creak", and a variant "whole creak" where an entire VʔV sequence is produced with creaky voice. Other variants include periods of breathy or regular modal voice or modal voice with reduced intensity. A significant effect of position in word is evident with more fully occluded variants in word-initial versus medial positions suggesting important positional variation. In a later study of the same corpus, Davidson and Parker Jones (2023) collapse variants into four main categories, fully occluded, creaky voice, breathy voice, and modal voice in order to investigate metrical context as well as word position on phonetic implementation of /ʔ/. They find that while more fully occluded variants occur in word initial versus medial position, main stress in itself does not appear to be a strong predictor across the board. Rather word-internal metrical boundaries seem to be a better predictor of where a more occluded variant will occur.

Given the general cross-linguistic observations summarized above as well as the detailed analyses of Hawaiian glottal stop variation (Davidson 2021; Davidson & Parker Jones 2023), we present in the following sections an overview of some of our current experimental research on Tahitian investigating the phonetic implementation and context of glottal stop variation.

2. Glottal stop variation and vowel glottalization in Tahitian

As mentioned earlier, until relatively recently there had been no focused quantitative experimental study of glottal stop variation or vowel glottalization in Tahitian. The impressionistic observations of earlier descriptions of spoken Tahitian (e.g. Lemaître 1973; Coppenrath & Prévost 1975) suggests that there is indeed glottal stop variation. In order to explore this variation, a corpus of elicited record-

ings was collected in Papeete, 2018 (see Fletcher & Gregory 2024). In keeping with recommended field phonetic practice, a variety of genres were collected (controlled sentence frames, text reading, spontaneous elicited narratives). We discuss results of our analysis of the first two genres here. It is well known that the typical laboratory-phonology type tasks where words of interest are inserted in sentence frames are likely to be hyperarticulated because they are generally produced in narrow focused contexts (see Beckman 1996; Gordon 2014). Consequently, readings of a Tahitian translation of the fable *La Bise et le Soleil* 'The Northwind (Southeast wind) and the Sun' were analysed presuming that experimental task would affect distribution of variants given it is well known to be a factor in speech sound production. We recorded six female speakers who were all born in Tahiti and are bilingual speakers of Tahitian and Tahitian French. The materials come from two of the elicited reading tasks. The first set of materials (Experiment 1, henceforth EXP1) were designed to illustrate all contrastive stops and vowels and consisted of two disyllabic words written in Tahitian orthography inserted in a carrier phrase. An example of the EXP1 sentence frame in Tahitian orthography and a general English translation is shown in (1). Please note that orthographically in Tahitian, the glottal stop or *'eta* is indicated by an apostrophe (Fare Vāna'a 1999).

(1) *I roto i te reo tahiti, e parauhia **pata** e'ere **pa'a**.*
'In Tahitian we say **scorpion** and not **bark**.'

Speakers produced 3–4 repetitions of each token in the sentence frame experiment and at least 3 or 4 repetitions of the read text task (our second set of materials; henceforth BES). The audio files were forced aligned using a version of the general WebMaus protocol (Kisler et al. 2017) that we adapted for French Polynesian language inventories, and vowel and consonant boundaries were subsequently checked and manually corrected using visual cues in the waveform and spectrogram. Stress was also annotated following the stress rules outlined by Bickmore (1995) for main stress (word accent) assignment summarized earlier. Most words have initial main stress in the sentence frame task unless they contain a second long vowel or diphthong which then attracted main stress. In all cases in the sentence frame task words are in focal position and therefore stressed syllables also carry a post-lexical pitch accent. For sake of comparability across the two tasks, accented syllables in all of the following analyses and discussion refers to lexically stressed syllables that also carry a post-lexical pitch accent.

Given our primary interest is glottal stop *variation*, different types of glottal stop variants were also annotated based on auditory and visual inspection of waveforms and spectrograms adapting the fine-grained annotation procedures in Davidson (2021) but for the purpose of this paper we have collapsed the variants into the four main categories used by Davidson and Parker Jones (2023) for Hawai-

ian. The key variants are as follows: full glottal closure (including at least 50% voicelessness), creaky voiced, breathy voiced, and modal voice. Creaky variants were noted if /ʔ/ was produced with aperiodic intermittent creaky voice with no F0, or with periodic but "constricted" voicing with low or irregular F0, and no evidence of full/partial voiceless glottal closure. Breathy voiced variants were noted if there was a high level of spectral noise and energy suggesting partially open vocal folds, and modal voiced variants were registered if there was no evidence of vocal fold adjustment associated with either creaky or breathy phonation.

Figure 1 shows speech waveforms and spectrograms of an inter-vocalic fully-occluded glottal stop variant and a creaky voiced variant of /ʔ/ (represented in the illustration as *q* for purely practical purposes in the phoneme tier) from the read text task of our Tahitian corpus for the word *mara'amu* /maraˈʔamu/ 'southeast wind'. The two examples are produced by two different speakers. There is a clear period of complete voiceless closure in the first spectrogram and speech waveform with one or two glottal pulses evident during the closure phase.

An example of a creaky voice variant is shown in the second spectrogram of Figure 1. The period of /ʔ/ constriction is shorter than the first variant and the absence of any F0 signal was apparent during the consonant interval. This closely resembles aperiodic creaky voice with intermittent rates of vocal fold vibration (wider spaced glottal pulses) compared to neighbouring vowels (after Keating et al. 2015). A small number of glottal pulses are also visible in the vowel following /ʔ/ in both instances suggesting vowel glottalization. It is well known that segmentation of these kinds of variants can be challenging and the reader is referred to Garellek et al. (2023) for further discussion.

A range of acoustic parameters were also extracted for quantitative analysis of voice quality following Garellek (2022). VoiceSauce (Shue et al. 2011) was used to extract harmonics-to-noise ratio < 1500 Hz (HNR15) which is a measure of inharmonic noise below 1500 Hz. Creaky phonation is generally associated with lower HNR values compared to modal phonation which typically has higher levels of periodicity. Other parameters were also extracted including Strength of Excitation (SoE) a measure of voice source energy/intensity, a spectral tilt measure corrected for vowel (H1*–H2*) with low values in both cases expected for creaky phonation, and finally voicing proportion calculated using Praat's Voice Report protocol (Boersma & Weenink 1992–2025) that gives a good indication of periodicity. It was assumed that higher voicing proportion values will suggest more creaky voice versus voiceless variants. These acoustic parameters were extracted for all labelled /ʔ/ and /p,t/ as well as accented and unaccented flanking vowels across the two experimental tasks. A summary of the number of consonants analysed is presented in Table 2. In the first instance, we present an overview of major variants, and then present some key results of the acoustic phonetic analyses of the corpus.

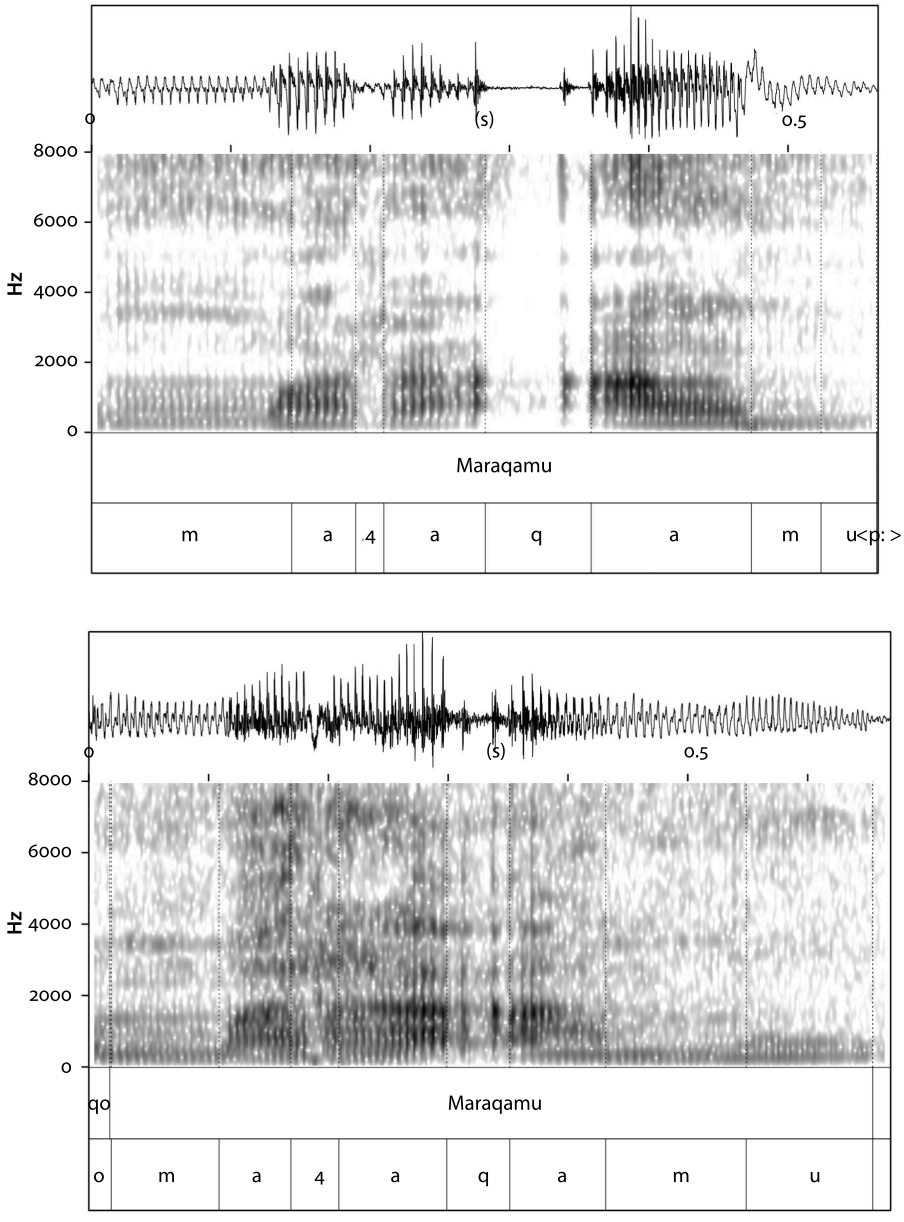


Figure 1. Two spectrograms showing two phonetic implementations of an intervocalic Tahitian glottal stop /ʔ/ for the word *mara'amu* /maɾaˈʔamu/ 'southeast wind'. /ʔ/ is represented by *q* in the phoneme label tier; /ɾ/ by *4*. The first shows a variant with full glottal closure and the second shows a creaky voiced "mid-creak" variant

Table 2. Number and distribution of /ʔ/ and combined oral stops /p/ /t/ according to experimental task and word position

	EXP1		BES	
	Initial	**Medial**	**Initial**	**Medial**
/ʔ/	377	701	302	306
/p/ /t/	1502	1057	484	209

2.1 Distribution of glottal variants

Figure 2 shows the distribution of major glottal stop variants plotted separately for experimental task and according to where they occur in a prosodic word (initial versus medial). In both experimental tasks, the dominant variant of /ʔ/ has full glottal closure with 77% of all variants in the sentence frame task and 59% in the BES corpus, so experimental task is an important factor determining the occurrence of canonical glottal stops as one would expect. This difference was confirmed statistically using a multinomial logistic regression analysis of variant distribution. A higher proportion of creaky voiced (including intervocalic mid-creak and whole creak variants with periodic low F0 creaky voice produced across a VʔV sequence) and other variants (breathy and modal) are realized in the BES task compared to the more controlled task although the latter are far less frequent overall ($z = 124.59$, $p < 0.0001$). In the sentence frame task, almost all word-initial variants are canonical voiceless glottal stops compared to 65% in word-medial position. In the BES task, around 66% of word initial glottal stops are produced with full glottal closure with 50% of medial /ʔ/ realized as canonical glottal stops ($z = 101.37$, $p < 0.0001$). We fully expected there to be an effect of accent, and there is across the corpus with more voiceless glottal stops realized in accented syllable onset positions ($z = 29.49$, $p < 0.0001$).

While creaky voiced variants are produced in medial contexts in both tasks, it should be noted that more lenited vowel-like whole-creak variants and modal voiced variants are observed in medial position in the BES corpus compared to the more controlled task. Only a handful of these variants are produced by the same speakers in EXP1. There are more identical vowel VʔV sequences in the BES corpus which is a prime location for either whole-creak or modal voice realizations according to Lemaître (1973). Moreover, most medial modal voiced variants in the sentence frame task are found in the minimal pair *roʔa* /roʔa/ 'heartwood' and *rōʔā* /roːʔaː/ 'shrub used for fishing lines' which is produced by two speakers without any glottalization. In general, there are only a handful of modal or breathy variants of /ʔ/ produced by speakers compared to the two major variants (full closure, creaky voice) particularly in the sentence frame task so these were excluded from the acoustic analyses presented in the following sections.

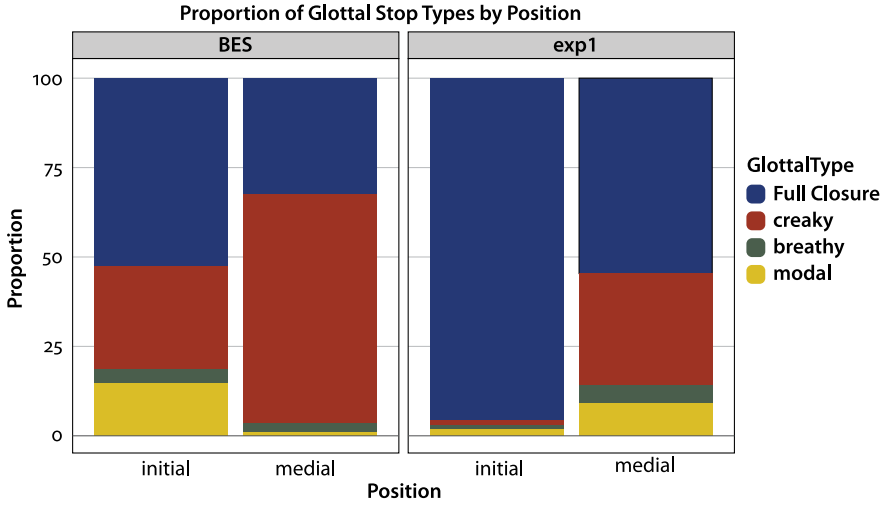


Figure 2. Proportional distribution of glottal variants of /ʔ/ across the corpus, grouped according to prosodic word position (initial, medial) and experimental task (sentence frame EXP1, and read text BES)

2.2 Acoustic duration of /ʔ/ and glottal variants

In other languages with phonemic glottal stop (e.g. Arapaho; Whalen et al. 2023), /ʔ/ is phonetically shorter than other stops. This is also the case in our Tahitian corpus. Figure 3 shows a boxplot summarizing the simplified distribution of acoustic duration values for all phonemic glottal stops /ʔ/ compared to the oral stops /p,t/ plotted according to whether they are part of an accented syllable onset or not (recall that accented here refers to post-lexical accentuation). The duration data were also submitted to a linear mixed effects model with fixed factors consonant type (oral versus glottal), prominence (accented versus unaccented), prosodic word position (initial versus medial), and experimental task (EXP1 versus BES). As can be seen in Figure 3, oral stops are significantly longer than glottal stops in general by around 60 ms ($F = 30.72$, $p < 0.0001$) and accent also has a similar effect overall ($F = 12.86$, $p < 0.0001$). Previous work has found significant effects of post-lexical accentuation on overall syllable duration (Fletcher 2021) so the latter result is not that surprising. All three stops are generally longer in the controlled experimental task compared to the BES task ($F = 12.11$, $p < 0.0001$).

A closer inspection of the duration patterns of the two main /ʔ/ variants (Figure 4) shows that full stop realizations are clearly longer than creaky voiced variants across both tasks ($F = 41.86$ $p < 0.0001$), with differences clearest in the sentence frame task. These findings are reminiscent of those reported for Arapaho, for example (e.g. Whalen et al. 2023) and for singleton glottal stops versus oral stops in Maltese (Mitterer 2018).

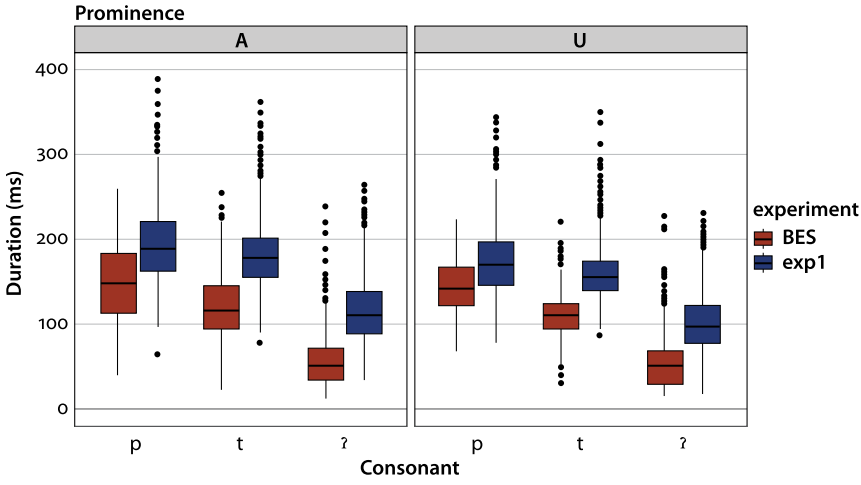


Figure 3. Glottal stop /ʔ/ and oral stop /p,t/ duration plotted by experimental task and accent status (A for accented, U for unaccented)

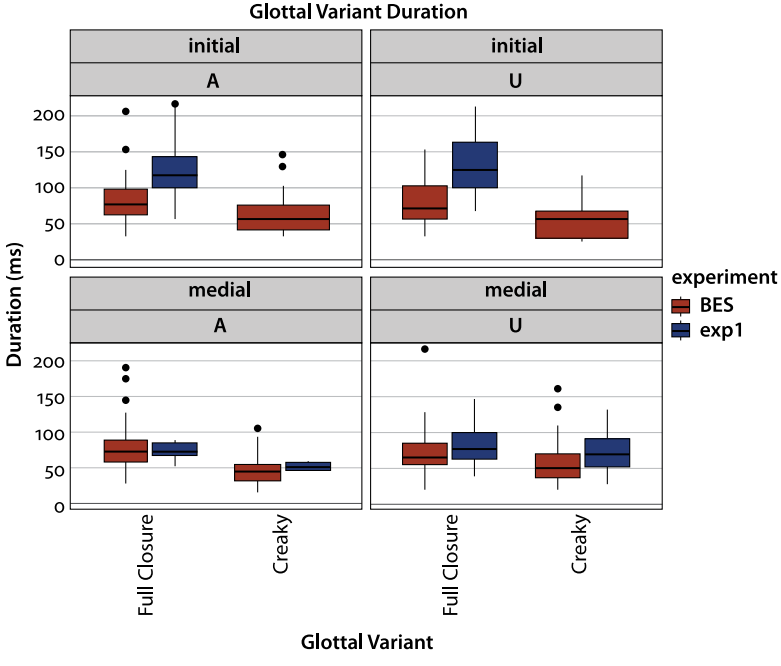


Figure 4. Durational variation of the two main /ʔ/ variants plotted by experimental task, prosodic word position, and accent status (A for accented, U for unaccented)

There are fewer non-full glottal closure variants in prosodic word-initial position in the sentence frame task with only three creaky voiced variants produced in accented word initial position, so these have been removed from Figure 4 for ease of presentation. It should be noted that these word-initial creaky voiced variants are only slightly shorter than full closure variants in this prosodic position. In fact, the majority of creaky voice variants typically exhibit intermittent aperiodic creaky voice of relatively low periodicity and amplitude, and are very much at the "consonantal" end of the creaky voice continuum, with more instances of vocal fry-type creak in medial positions when flanking vowels are identical. This is highly reminiscent of Davidson's (2021) "whole creak" variant in Hawaiian, where the entire V?V sequence is realized as a period of creaky phonation. There are small duration differences between the two main variants in prosodic word-medial position across the two tasks (16 ms). Based on this analysis and categorization of variants outlined previously, prosodic word-initial position in Tahitian appears to be a prosodically prominent context, adding another language to the recent largescale study of initial consonant lengthening across the 51 languages of the DORECO corpus by Blum et al. (2024). Note however that accentual status is less consistent, and this could well be because in the sentence frame task most word-initial tokens are also accented syllable onsets. In the two datasets explored here, task and prosodic word position are more general predictors of durational variation of /ʔ/.

2.3 Voice quality measures of /ʔ/ and neighbouring vowels

In this section we present key acoustic analyses of spectral noise (Strength of Excitation—SOE) and voicing proportion (a measure of periodicity) for /ʔ/ regardless of variant. The voicing fraction data were submitted to statistical investigation and a linear mixed effects model was run with fixed factors prominence, word position, and task. The strength of excitation signal data was submitted to a generalised additive mixed effects model with the same fixed effects. Given more creaky voiced variants are observed in medial contexts and in the read text task we expected higher levels of source noise and voicing fraction here compared to word initial positions where there are more voiceless or partially voiceless stop-like variants. Moreover, given more of the latter are found in the sentence frame task, we fully expected this variation would also show up in our measures. Inspection of Figure 5 shows this is mostly the case with /ʔ/ associated with less noise (lower dB levels) and lower degrees of periodicity (proportion of voicing) in the sentence frame task (EXP1) compared to the read text task (SoE: parametric $t = -25.046$, $p < 0.0001$; Voicing Fraction: $F = 6.01$, $p < 0.01$). Prominence also has a significant effect on voicing fraction ($F = 7.34$, $p < 0.01$). /ʔ/ in unaccented syllable onset positions has higher and

more variable levels of periodicity (i.e. more creaky phonation overall in the read text task (BES) which is somewhat interesting but this also likely reflects the more diverse lexical structure of the materials with many glottal stop-initial grammatical particles and so forth compared to the sentence frame task.

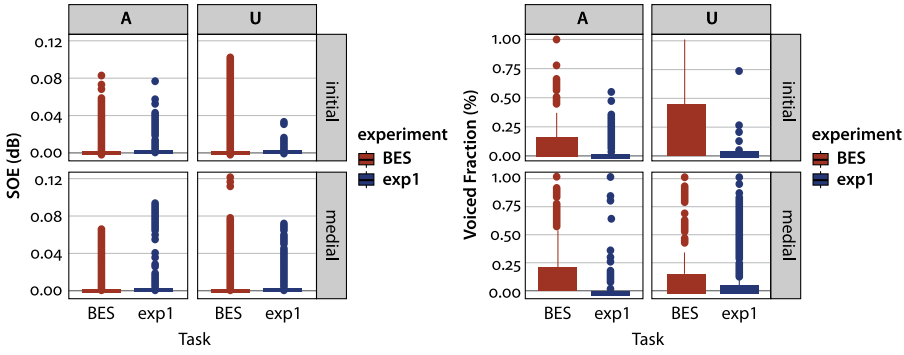


Figure 5. Strength of Excitation (SoE) and Voicing fraction values in Tahitian /ʔ/ plotted by position in prosodic word, accent status (A for accented, U for unaccented) and task

Turning now to neighbouring vowels, it was clear in our initial finer-grained annotation of the corpus (Fletcher & Gregory 2024), that there is a degree of variation in the extent of vowel glottalization. Many variants were originally labelled V1 creak or V2 creak (after Davidson 2021) when a substantial portion of the vowel preceding or following /ʔ/ shows creaky phonation in the waveform/spectrogram so we fully expected our acoustic measures to reflect this. Indeed, there is strong glottalization based on two further acoustic measures of voiced quality: HNR (Harmonics-to-Noise ratio), and spectral tilt (H1*–H2*). We also see a fairly complex interplay of accentual prominence and prosodic word position in these data. A total of 5,466 preceding and following vowels (flanking either /p,t/ or /ʔ/ were analysed and submitted to a generalized additive mixed effects model with fixed factors consonant type, task, prosodic word position, and prominence. Figure 6 plots the time course of loess-smoothed spectral noise (HNR15) and spectral tilt (H1*–H2*) signals for vowels flanking glottal stops (solid lines) and oral stops (dotted lines).

For ease of presentation, the parameters are presented across the two tasks given similar trends were observed in each case. Medial unaccented vowels in glottal environments have lower overall spectral noise (HNR) values than in oral stop environments (Parametric: $t = 37.67, p < 0.0001$; Smooth term: $F = 45.62, p < 0.0001$) suggesting lower levels of periodicity and more glottalization compared to oral stop contexts. There are also significant differences in spectral tilt values according to consonant type (Parametric: $t = 7.92$; Smooth term: $F = 66.76, p < 0.0001$). For spec-

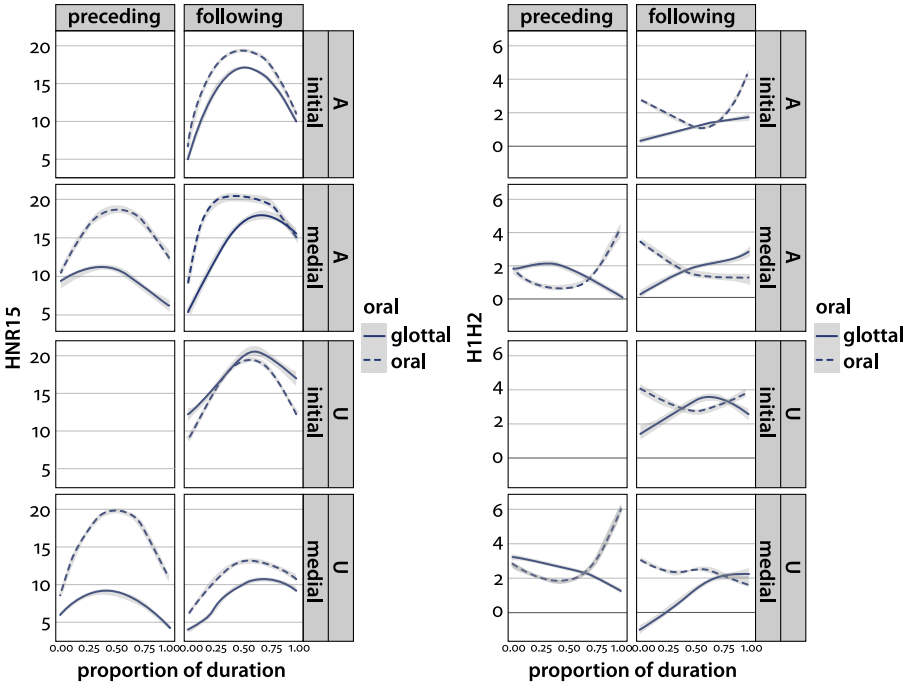


Figure 6. Time course of Loess-smoothed Harmonics-to-Noise ratio (left) and Spectral Tilt (H1*–H2*) (right) for vowels either side of /ʔ/ and oral stops, plotted according to stress (A for accented/stressed, U for unaccented/unstressed)

tral tilt, the two signals (oral versus glottal) significantly diverge in the last 25% of preceding vowels, and during the initial 25–50% of following vowels with *lower* values for glottal versus oral stop environments as one would expect with clear carryover and anticipatory creaky voice in the vicinity of /ʔ/. Differences are more extensive, i.e. there are longer periods of glottalization for vowels that follow /ʔ/ (i.e. those that are part of the same syllable), thus confirming Lemâitre's (1973) general observation that more glottalization is observed in vowels following /ʔ/. The highest degree of carryover coarticulation is observed in prosodic word-medial unaccented vowels with glottal stop onsets where we see more creaky voiced variants, particularly in the read text task.

There is a potential conundrum with regard to accented word initial contexts, however, where more canonical stop-like /ʔ/ variants are produced, particularly in the sentence frame task. We might have predicted that if initial accented syllables are prime locations for gestural strengthening (which is certainly the case for /ʔ/ implementation i.e. this is where most full-closure variants are observed), we would see spectral tilt values that are *higher* than in unaccented medial contexts,

for example (compare the top right and bottom right panels of Figure 6) i.e. *less* glottalization in this context than we actually observe. Instead, we see *lower* overall spectral tilt values in the initial 25% or so of vowels that are in prosodic word-initial and accented position (top right panel of Figure 6) although not to the same temporal extent as in medial unaccented following vowel contexts (bottom right panel). This could indicate greater vocal fold constriction but these vowels also have *higher* HNR values and *higher* voiced fraction suggesting more periodicity in the initial 25–50% compared to vowels in medial unaccented contexts. It could be that given the prosodic salience of accented initial syllable context and the dominance of full closure (partially) voiceless variants in this location, this is sufficient to cue this important phonemic contrast. It is well known there is no single acoustic correlate for creaky voice, and languages differ in their patterns of implementation of /ʔ/ (e.g. Davidson 2021; Keating et al. 2023; Di Canio 2021). There are also potential interactions with prosodic factors as appears to be the case in our data. In oral stop environments we also see *higher* HNR but *higher* spectral tilt values (i.e. less glottalization) in accented syllables so we cannot assume the difference in spectral tilt values in accented word-initial syllables is due to an across the board effect of accentuation. It is most likely that the voice quality measures are picking up on levels of gradience initially factored into our earlier finer-grained annotations, with more lenited variants (lower spectral tilt, lower HNR, more variable and lower levels of voicing fraction) showing up in prosodic word-medial unaccented environments across the two tasks. Certainly, descriptive studies of glottal stop variation and glottalization in the closely related language Marquesan (Mutu & Teìkitutoua 2002) also suggest there are degrees of "creakiness" and similar kinds of gradience are observed across our corpus within the broader creaky voiced category adopted in this chapter. In future work we may reanalyse the creaky voice category according to the different kinds of creaky phonation reviewed in Keating et al. (2015). This kind of variation is also not uncommon cross-linguistically and the reader is also referred to Esling and Harris (2005) for further discussion of this issue.

3. Discussion

In this chapter, we hope to have shown that exploring the fine phonetic detail of languages with small phoneme inventories gives additional insight into the complex interplay between the implementation of key phonemic categories and prosodic factors in a language. Languages with a small stop series like Tahitian show patterns of variation that are of great typological interest, particularly in view of largescale crosslinguistic analyses of the glottal stop (Garellek et al. 2023).

There is a degree of expected variation the realization of /ʔ/ in Tahitian, and the range of glottal variants observed in our Tahitian corpus echoes findings for other languages of the world which also have phonemic glottal stops (Garellek et al. 2023; Whalen et al. 2023), many of which have more complex consonant inventories. Crucially however, we see a different pattern of implementation to that noted for the Eastern Polynesian language, Hawaiian, that has received a lot of attention in recent years (e.g. Davidson 2021; Davidson & Parker Jones 2023). The proportion of canonical glottal stops in our Tahitian corpus is far higher with 59% realized as fully occluded glottal stops, compared to 7% reported for Hawaiian. It should be made clear that the corpus analysed by Davidson (2021) and Davidson and Parker Jones (2023) consists of radio interviews so further work is needed on our spontaneous speech corpus to explore similar kinds of naturalistic speech for Tahitian to see if we get similarly low levels of canonical glottal stops.

The distribution of /ʔ/ variants differ across speaking task in our corpus so task is clearly a factor. More gesturally-strengthened canonical glottal stops are produced in sentence frame contexts as a proportion of all glottal variants, and we suspect this also reflects the influence of information structure in speech segment realization in classic sentence frame paradigms with the token of interest realized under narrow focus. Our corpus also highlights phonetic gradience even within the category of canonical glottal stops with many /ʔ/ variants in the *La Bise et le Soleil* corpus, whilst showing a period of voicelessness, often also show a period of incomplete closure reinforcing the highly gradient nature of glottal stop production as commonly stated by Garellek (2014, 2022) among others. Moreover, the creaky voice category in this study also shows a degree of variation in its phonetic implementation as noted in the previous section. Nevertheless, what is particularly striking in our corpus is that glottal stops in Tahitian tend to be more "consonantal" and stop-like rather than "vocalic". Even many creaky voice realizations tend to be realized as low amplitude periods of intermittent aperiodic voice compared to more vocalic periodic creaky voice realizations (i.e. more reminiscent of vocal fry). Whilst not explored in any detail in this paper, the oral stops are also short lag in our corpus (e.g. see Fletcher et al. 2023) and show very little variation in their realization, i.e. they are never lenited or voiced unlike the kinds of variation you see in other single stop series languages, thinking here in particular of many Australian indigenous languages (e.g. Butcher 1995).

Position in word, and to a certain extent, level of accentual prominence are also key factors in determining whether canonical glottal stops are realized (particularly in *La Bise et le Soleil* readings) although prosodic word position or genre are more influential in this regard. More lenited creaky voiced variants are observed word-medially as one would predict particularly in unaccented medial contexts in the read text corpus. Higher proportions of true glottal stops are

also observed in initial and medial position in the controlled task. Acoustically, full closure variants (i.e. voiceless or partially voiceless /ʔ/) tend to be longer than other variants, with relatively low intensity and voicing fraction values (i.e. less periodicity) suggesting gestural strengthening, particularly in prosodic word-initial accented syllable onsets. Our findings also support some of the earlier observations by Lemaître (1973) and Coppenrath and Prévost (1975) who mention coarticulatory glottalization in their overviews of the phonology of Tahitian, but it is clear that prosodic context is also a key factor here with a different type of phonetic implementation evident in positions of prosodic salience in our corpus.

Our work takes place at a time when there has been an upsurge of large-scale cross-linguistic phonetic analyses of field corpora as mentioned in the introduction to this chapter. This means that the time is ripe to review typically held views about phonetic and phonological typology. We are in complete agreement with Himmelmann (2023) when he affirms that advances in cross-linguistic phonetic and phonological typology can only be made through a more nuanced empirical approach. Our case study of Tahitian shows that an exploration of the fine phonetic detail of deceptively simple phonological contrasts reveals important patterns of phonetic implementation with variation mitigated by prosodic context and spoken genre.

Funding

Aspects of this research were supported by the Australian Research Council Centre of Excellence for the Dynamics of Language CE1401000 and a mission de recherche award from the Université de la Polynésie Française awarded to the first author.

Acknowledgements

Heartfelt thanks are expressed to our speakers, and to Jacques Vernaudon and Mirose Paia for assistance with subject recruitment in Papeete and material development. Eleanor Lewis and Ben Volchok also assisted with earlier material preparation and annotation of the corpus. Our thanks also to the editors and to two anonymous reviewers.

References

Beckman, Mary E. 1996. "The Parsing of Prosody." *Language and Cognitive Processes* 11: 17–68.

Bickmore, Lee S. 1995. "Revizing and Formalizing the Tahitian Stress Placement Algorithm." *Oceanic Linguistics* 34 (2): 410–442.

Blum, Frederic, Ludger Paschen, Robert Forkel, Susanne Fuchs, and Frank Seifart. 2024. "Consonant Lengthening Marks the Beginning of Words Across a Diverse Sample of Languages." *Nature Human Behavior* 8: 2127–2138.

Bodin, Vonnick. 2006. *Tahiti: la langue et la société*. Papeete: Éditions 'Ura.

Boersma, Paul, and David Weeninck. 1992–2025. "Praat: Doing phonetics by computer." Retrieved 5 May 2025.

Boula de Mareüil, Philippe, Frédéric Vernier, Gilles Adda, Albert Rilliard, and Jacques Vernaudon. 2019. "A Speaking Atlas of Indigenous Languages of France and Its Overseas Territories." *Proceedings of International Conference Language Technologies for All (LT4All)*: 155–159.

Butcher, Andrew. 1995. "The Phonetics of Neutralisation: The Case of Australian Coronals." In *Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor*, ed. by Jack Windsor-Lewis, 10–38. New York: Routledge.

Charpentier, Jean-Michel, and Alexandre François. 2015. *Atlas linguistique de la Polynésie française*. Berlin: De Gruter Mouton.

Cho, Taehong. 2016. "Prosodic Boundary Strengthening in the Phonetics–Prosody Interface." *Language and Linguistics Compass* 10: 120–141.

Chodroff, Eleanor. 2025. "Phonetic Universals." *Annual Review of Linguistics* 11: 251–273.

Coppenrath, Hubert, and Paul Prévost. 1975. *Grammaire approfondie de la langue tahitienne (ancienne et moderne)*. Papeete: Librairie Pureora.

Davidson, Lisa. 2021. "Effects of Word Position and Flanking Vowel on the Implementation of Glottal Stop: Evidence from Hawaiian." *Journal of Phonetics* 88: 1–14.

Davidson, Lisa, and Oiwi Parker Jones. 2023. "Word Level Prosodic and Metrical Influences on Hawaiian Glottal Stop Realization." *Phonetica* 80 (3–4): 225–258.

Davies, John. 1851. *A Grammar of the Tahitian Dialect of the Polynesian Language*. Tahiti: The Mission Press.

Di Canio, Christian T. 2021. "Phonetic variation in the production of glottal stops and glottalization." Presentation, Haskins Laboratories. Retrieved 19 September 2025 from https://www.acsu.buffalo.edu/~cdicanio/talks.html

Di Canio, Christian T., and Ryan Bennett. 2020. "Prosody in Mesoamerican Languages". In *The Oxford Handbook of Language Prosody*, ed. by Carlos Gussenhoven, and Aoju Chen, 408–427. Oxford: Oxford University Press.

Esling, John H., Scott R. Moisik, Allison Benner, and Lise Crevier-Buchman. 2019. *Voice Quality: The Laryngeal Articulator Model*. Cambridge: Cambridge University Press.

Esling, John H., and Jimmy G. Harris. 2005. "States of the Glottis: An Articulatory Phonetic Model Based on Laryngoscopic Observations." In *A figure of speech: A Festschrift for* John Laver, ed. by William J. Hardcastle, and Janet Mackenzie Beck, 347-383. Mahwah, NJ: Lawrence Erlbaum Associates.

Fare Vāna'a. 1999. *Dictionnaire tahitien–français*. Tahiti: Academie Tahitienne.

Fare Vāna'a. 2009. *Grammaire de la langue tahitienne*. Tahiti: Academie Tahitienne.

Fletcher, Janet. 2021. "Intonational Phonetics and Register Variation in Two Oceanic Languages." Presentation, Symposium Forms and Representations in Phonology: In Honour of Bob Ladd. Barcelona, 23 June 2021. https://pape2021.upf.edu/session/phonology/

Fletcher, Janet, Adele Gregory, and Ben Volchok. 2023. "Acoustic Correlates of Oral and Glottal Stops in Tahitian." *Journal of the Acoustical Society of America* 154: A76.

Fletcher, Janet, and Adele Gregory. 2024. "Speech Task and Prosodic Context Influence Glottal Stop Variation in Tahitian." In *Proceedings of the 19th Australasian International Conference on Speech Science and Technology*, ed. by Olga. Maxwell, and Rikke Bundgaard-Nielsen, 162–166. Melbourne: Australasian Speech Science and Technology Association.

Garellek, Marc. 2014. "Voice Quality Strengthening and Glottalization." *Journal of Phonetics* 45 (1): 106–113.

Garellek, Marc. 2022. "Theoretical Achievements of Phonetics in the 21st Century: Phonetics of Voice Quality." *Journal of Phonetics* 94: 101155.

Garellek, Marc, Yuan Chai, Yaqian Huang, and Maxine van Doren. 2023. "Voicing of Glottal Consonants and Non-Modal Vowels." *Journal of the International Phonetic Association* 53 (2): 305–332.

Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel. 2006. *Essentials of Language Documentation*. Berlin: De Gruyter Mouton.

Gordon, Matthew K. 2014. "Disentangling Stress and Pitch-Accent: A Typology of Prominence at Different Prosodic Levels." In *Word Stress: Theoretical and Typological Issues*, ed. by Harry van der Hulst, 83–118. Cambridge: Cambridge University Press.

Gordon, Matthew K. 2016. *Phonological Typology*. Oxford: Oxford University Press.

Gussenhoven, Carlos, and Aoju Chen (eds). 2020. *The Oxford Handbook of Language Prosody*. Oxford: Oxford University Press.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36: 161–195.

Himmelmann, Nikolaus P. 2006. "Prosody in Language Documentation." In *Essentials of Language Documentation*. ed. by Jost Gippert, Nikolaus. P. Himmelmann, and Ulrike Mosel, 163–182, Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2023. "On the Comparability of Prosodic Categories: Why 'Stress' is Difficult." *Linguistic Typology* 27 (2): 341–361.

Himmelmann, Nikolaus P., and D. Robert Ladd. 2008. "Prosodic Description: An Introduction for Fieldworkers." *Language Documentation and Conservation* 2: 244–274. http://hdl.handle.net/10125/4345

Jaussen, Tepano. 1886. *Grammaire et dictionnaire de la language maorie: Dialecte tahitien*. Braine-le-Comte: Zech et fils.

Jun, Sun-Ah. 2005. "Prosodic Typology." In *Prosodic Typology: The Phonology of Intonation and Phrasing*, ed. by Sun-Ah Jun, 430–458.

Jun, Sun-Ah, and Janet Fletcher. 2014. "Methodology of Studying Intontation: From Data Collection to Data Analysis." In *Prosodic Typology II: New Developments in the Phonology of Intonation and Phrasing*, ed. by Sun-Ah Jun, 493–519. Oxford: Oxford University Press.

Kaland, Constantijn. 2023. "Contour Clustering: A Field-Data-Driven Approach for Documenting and Analysing Prototypical F0 Contours." *Journal of the International Phonetic Association* 53 (1): 159–188.

Keating, Patricia A., Taehong Cho, Cécile Fougeron, and Chai-Shune Hsu. 2003. "Domain-initial Strengthening in Four Languages." *Papers in Laboratory Phonology* 6: 145–163.

Keating, Patricia A., Marc Garellek, and Jody Kreiman. 2015. "Acoustic Properties of Different Kinds of Creaky Voice." *Proceedings of the 18th International Congress of Phonetic Sciences* 821: 1–5.

Keating, Patricia A., Jianjing Kuang, Marc Garellek, Christina M. Esposito, and Sameer ud Dowla Khan. 2023. "A Cross-Language Acoustic Space for Vocalic Phonation Distinctions." *Language* 99 (2): 351–389.

Kisler, Thomas, Uwe Reichel, and Florian Schiel. 2017. "Multilingual Processing of Speech Via Web Services". *Computer, Speech, and Language* 45: 326–347.

Kohler, K. J. 1994. "Glottal Stops and Glottalization in German." *Phonetica* 51 (1–3): 38–51.

Ladefoged, Peter, and Ian Maddieson. 1996. *The Sounds of the World's Languages*. Oxford: Wiley-Blackwell.

Lazard, Gilbert, and Louise Pelzer. 1991. "Predicates in Tahitian." *Oceanic Linguistics* 30 (1): 1–31.

Lemaître, Yves. 1973. *Lexique du tahitien contemporain*. Paris: ORSTOM.

Lynch, John. 1998. *Pacific Languages: An Introduction*. Honolulu: University of Hawai'i Press.

Maddieson, Ian. 2001. "Phonetic Fieldwork." In *Linguistic Fieldwork*, ed. by Paul Newman, and Martha Ratliff, 211–229. Cambridge: Cambridge University Press.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. "Montreal Forced Aligner", vo.9.0. Retrieved 17 January 2017 from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

Mitterer, Holger. 2018. "Not All Geminates are Made Equal: Evidence From Maltese Glottal Consonants." *Journal of Phonetics* 66: 28–44.

Mutu, Margaret, and Ben Teìkitutoua. 2002. *Ùa Pou — Aspects of a Marquesan Dialect*. Canberra: Pacific Linguistics.

Nicholas, Sally Akevai, and Rolando Coto-Solano. 2019. "Glottal Variation, Teacher Training, and Language Revitalization in the Cook Islands." In *Proceedings of the 19th International Congress of Phonetic Sciences*, ed. by Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren, 3602–3606. Melbourne: Australasian Speech Science and Technology Association.

Paia, Mirose, and Jacques Vernaudon. 2002. "Le tahitien: plus de prestige, moins de locuteurs." *Hermès* 32–33: 395–402.

Parker Jones, ʻŌiwi. 2017. "Hawaiian." *Journal of the International Phonetic Association* 48: 1–13.

Paschen, Ludger, Susanne Fuchs, and Frank Seifart. 2022. "Final Lengthening and Vowel Length in 25 Languages." *Journal of Phonetics* 94: 101–179.

Perini, Antoine-Dominique. 1987. "La phonologie du tahitien et l'intégration des emprunts." *La Linguistique* 23 (2): 131–142.

Shue, Yen-Liang, Patricia A. Keating, Chad Vicenik, and Kristine M. Yu. 2011. "VoiceSauce: A Program for Voice Analysis." *Proceedings of ICPHS 17*: 1846–1849.

Tryon, D. T. 1970. *Conversational Tahitian: An Introduction to the Tahitian Language of French Polynesia*. Berkeley: University of California Press.

Vernaudon, Jacques. 2011. "Grammaticalization of Tahitian Mea 'Thing, Matter' into a Stative Aspect." In *Topics in Oceanic Morphosyntax*, ed. by Claire Moyse-Faurie, and Joachim Sabel, 319–340. Berlin: De Gruyter Mouton.

Vernaudon, Jacques, Nick Thieberger, Tamatoa Bambridge, and Takurua Parent. 2021. "Un nouveau souffle numérique pour les corpus en langues océaniennes." *Journal de la Société des Océanistes* 153: 323–336.

Vernaudon, Jacques, Alexandre François, Alejandrina Cristia, and Marie Saläun. 2025–2029. "The HéLiCéO Project: Understanding the Diversity of Languages and How they are transmitted in a multilingual context." Retrieved 30 July 2025 from https://www.cnrs.fr/en/ri2-project/heliceo

Walworth, Mary. 2020. "Eastern Polynesia", In *The Routledge Handbook of Language Contact*, ed. by Evangelia Adamou, and Yaron Matras, 1–17. London: Routledge.

Whalen, Doug H., and Joyce McDonough. 2015. "Taking the Laboratory Into the Field." *Annual Review of Linguistics*. 1: 395–415.

Whalen, Doug H., Christian T. Di Canio, and Rikker Dockum. 2022. "Phonetic Documentation in Three Collections: Topics and Evolution." *Journal of the International Phonetic Association*. 52 (1): 95–121.

Whalen, Doug H., Cristian T. Di Canio, and Wei-Rong Chen. 2023. "Variable Realization of the Arapaho Glottal Stop Despite Its Being Distinctive and Frequent." In *Proceedings of ICPHS 20*: 3276–3280.

Morphosyntax

CHAPTER 19

High animacy does not favour zero over pronouns – especially for objects

Geoffrey Haig & Stefan Schnell
University of Bamberg | University of Zurich

Since Givón (1983) and Ariel (1990), a large body of research continues to assume that zero anaphors reflect a higher degree of accessibility than overt pronominal forms. However, when the features of animacy or person are taken as its metric, existing case studies do not consistently confirm a correlation between high accessibility and zero form, and some report the opposite tendency. In this chapter we bring a broader cross-linguistic perspective to the question, leveraging data from a multilingual richly annotated spoken language corpus covering 16 typologically diverse languages. We identify a surprisingly robust trend in support of a reversal of expected accessibility-accounts: Overall, it is non-human referents rather than human, and third person rather than first and second person, which favour zero expression. Our findings provide cross-linguistic support for the constraint 'avoid non-human pronouns', found by Genetti and Crain (2003) for Nepali, and a tendency to overtly realise speech-act participants as pronouns. Furthermore, we identify a significant effect of syntactic role, such that these regularities are more clearly evident for direct objects than for transitive subjects. We explore possible explanations for these results, noting similarities between our findings and well-documented patterns of differential indexing, informativity and surprisal (Haspelmath 2021). More generally, these findings call for a re-assessment of the view that a single principle such as accessibility can account for referential choice among all form types. For the choice between reduced forms (pronouns and zero), distinct principles may be operative.

Keywords: referential choice, accessibility theory, person and animacy, grammatical relations, corpus-based typology

https://doi.org/10.1075/slcs.240.19hai
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

1. Introduction

Over the last 50 years, it has become increasingly evident that even core aspects of morphosyntax are highly sensitive to discourse context. Exploring the nature of the interdependence between discourse and grammar is a large and rapidly expanding field of linguistics, fuelled by advances in information technology, cognitive science, and language documentation. This has enabled more systematic studies within a usage-based paradigm of comparative linguistics and diachronic typology through investigations of corpus data. Nikolaus Himmelmann has contributed in multiple ways to this development, including ground-breaking work in the development of language documentation as a specific academic enterprise distinct from language description (Himmelmann 1998, 2012), but also on specific topics in grammar and discourse such as demonstrative use in discourse and the diachronic development of determiner-like elements (Himmelmann 1997, 1998), the impact of prosodic parsing on morphologisation and the "suffixing preference" (Himmelmann 2014), the interrelation of syntactic and prosodic structure more generally (Himmelmann 2022), and the development of conceptions of prominence as relevant in linguistics (Himmelmann & Primus 2015). In a fine-grained investigation of 'Undergoer Oriented Predicate'-constructions[1] in spoken Tagalog narratives, Himmelmann (1999) demonstrates that anaphoric subject pronouns are unexpectedly frequent in contexts where cross-linguistically, zero anaphora would be predicted, which Himmelmann suggests may be evidence of incipient subject agreement in Tagalog (see Section 5 for discussion). These, and many other contributions, have informed and inspired much current literature, including this chapter. As will become apparent, our research, likewise based on spoken narratives, also identifies contexts where pronouns are unexpectedly preferred over zeroes.

The research reported here is based on a richly-annotated multilingual corpus of spoken language, from 16 geographically, culturally and typologically diverse languages (see Section 3). Specifically, we focus on the choice among reduced forms of reference (pronouns versus zero anaphor), and the impact of animacy and person of referents, further contrasting between transitive subjects (A) and transitive objects (P). Existing literature on this topic stems from several distinct research paradigms: pronoun resolution in NLP and psycholinguistics (Stevenson et al. 2000; Arnold et al. 2000; Kehler & Rhode 2013; Kehler 2002, 2006; Patterson & Schumacher 2020); usage-based typological approaches to discourse and grammar (Givón 1983; Clancy 1980 on Japanese vs. English), variationist studies (often focused on subject realisation, cf. Torres Cacoullos & Travis 2019), explorations

1. In later literature loosely referred to as Undergoer Voice.

of the pro-drop parameter (Holmberg 2009; Rosenkvist 2009, 2018) and treatments of implicit reference as determined by socio-cultural parameters and considerations of 'discourse styles' (Bisang 2014, 2015; Stoll & Bickel 2009; Duranti 1994). The literature is correspondingly vast and varied, and we here focus on the research tradition on referential choice initiated by Chafe (1976) and continued by Ariel (1988, 1990), Gundel et al. (1993), and later Arnold (2008, 2010) on the correlation between referent accessibility and the choices that speakers exercise in selecting communicatively appropriate linguistic forms (full NPs, pronouns, zero etc.). In this chapter, we focus entirely on the choice between pronominal and zero forms, a decision we will justify in the course of the paper. For a detailed complementary investigation of referential choice between lexical and reduced forms see Schiborr (2023), and for choices between all three NP, pronominal and zero see Linders and Schnell (under revision).

The paper is organised as follows: Section 2.1 sets out the foundational claims of accessibility and related approaches to referential choice, taking Ariel's Accessibility Theory as the point of departure. Section 2.2 turns to the paradox of "reduced forms" (pronouns and zeroes, Kibrik 2011), citing literature which suggests a reversal of the predicted correlation between high accessibility and lighter form. We also consider differences between subject and object here. In Section 3 we outline the design of our study, including specific research questions and the materials used. In Section 4 we summarise our results, which are discussed in Section 5. We conclude the paper in Section 6.

2. Theoretical background

2.1 Accessibility and referential choice

Since the 1970s there exists growing consensus about the basic mechanism deployed by language users to keep track of the entities they talk about during communication. Essentially, producers choose forms of reference according to their estimates of their addressees' state of mind, reserving more informative and more verbose forms of reference to those entities that they deem otherwise hardly predictable (Clark & Murphy 1982; Tily & Piantadosi 2009; Kibrik 2011; Gibson et al. 2019). Accessibility Theory (Ariel 1988, 1990), and its offshoots, remains one of the most influential contributions to reference production within usage-based linguistics. The theory is an attempt to identify general principles that underly speakers' choices when selecting communicatively appropriate forms for referring to discourse referents (in the sense of Karttunen 1976). According to Ariel (2000: 204),

> […] when reference to mental entities [i.e. discourse referents, GH & SS] is made, the speaker chooses her referring expression according to how she assesses the accessibility of the specific referent for her addressee at the current stage of the discourse. The higher the mental accessibility, the higher the accessibility marker chosen.

The degree of accessibility of an antecedent is determined by both discourse-contextual features, such as syntactic function and topicality, as well as referent-inherent features such as animacy, see discussion in Section 5. The specific claim is that the array of possible form types available to speakers of any given language can be ordered hierarchically from higher to lower accessibility markers.

In earlier literature, a distinction was introduced between "salience", which referred to properties of the antecedent such as animacy and syntactic role, and the "accessibility" of anaphoric mentions of that antecedent (Ariel 2000: 205–206). The salience of the antecedent thus contributes to the accessibility of the forms under investigation: a human antecedent is more salient than a non-human antecedent, and is thus predicted to trigger higher accessibility of the anaphor, and so on. In more recent literature, however, the distinction blurs, or overlaps with conceptually related terms such as Chafe's (1987) "activation", Gundel et al.'s (1993) "givenness", and more recently "prominence". For the purposes of this study, we will maintain the loose usage of the term "accessibility", as used in the accessibility scale in (1), while acknowledging that this is conceptually an over-simplification; see Boswijk and Coler (2020) on a scientific definition of "salience" within linguistics, and von Heusinger and Schumacher (2019) for a clarification of "prominence" versus related concepts.[2] We also note that the Accessibility Scale has clear precursors in Givón's (1983: 17) "topic-continuity scale", see e.g. Huang (2000: 303–305) for discussion.

According to Accessibility Theory, a single unifying principle is operative across the board, motivating not only the choice of a 'full description' versus any reduced form of reference, or a proximal demonstrative over a distal demonstrative, but also choices between different kinds of reduced forms, such as zero over a pronoun (cf. Ariel 1990: 73–76, cf. Gundel et al. 1993 for explication of analogous claims). A further, though logically independent, principle underlying scales such as (1), is that the highest accessibility markers have phonologically the least substance: pronouns (strong and weak), agreement morphology, or zero anaphor are assumed to be the forms of choice for discourse referents with maximal accessibility: "Speakers choose pronouns or zero forms for salient discourse

2. We also point out that the term "accessibility" is used both as a proxy for the speaker's assessment of the cognitive status of the referent in the mind of the listener, but also as a characterisation of linguistic forms deployed by the speaker ("acccessiblity markers").

referents (e.g., Arnold 1998; Kehler et al. 2008)." (von Heusinger & Schumacher 2019: 125). This links back to the more general idea that languages are maximally efficient, deploying forms with less substance for highly predictable — here: accessible — information (Levy & Jäger 2007). Tily and Piantadosi's (2007) experimental study does indeed suggest fairly minute alignment between producers and recipients with the latter making exactly correct predictions on intended referents based on the deployment of more or less informative forms of reference. However, their study was confined to 'pronominalisation', i.e. the choice between a fully descriptive and informative NP and a less informative pronoun. Matters become less clear-cut once choices between high-accessibility markers are considered vis-à-vis prominence-relevant features such as humanness and person. Moreover, the choice between such high-accessibility markers seems to play out differently across syntactic functions, as we will show in Section 2.3.

(1) Accessibility Scale, reduced forms shaded (adapted from Ariel 2000: 205)

	Accessibility	
a.	Full name + modifier	Low
b.	Full ("namy") name	
c.	Long definite description	
d.	Short definite description	
e.	Last name	
f.	First name	
g.	Distal demonstrative + modifier	
h.	Proximal demonstrative + modifier	
i.	Distal demonstrative (+ NP)	
j.	Proximal demonstrative (+ NP)	
k.	Stressed pronoun + gesture	
l.	Stressed pronoun	
m.	Unstressed pronouns	
n.	Reduced/cliticised pronouns	
o.	Rich agreement markers	
p.	Poor agreement markers	
q.	Reflexives	
r.	Zero	High

As we are reporting here on a cross-linguistic study into the validity of Accessibility Theory, it is worth pointing out that the theory is designed to be universally applicable to all human languages. While languages obviously differ in terms of their inventories of referential forms, the crucial claim here is that these forms

are ranked in terms of accessibility *in each language*, so that degree of accessibility completely controls their usage. For instance, while one language may possess demonstrative pronouns, 3rd person pronouns, both full and clitic ones as well as the possibility of zero reference, another language may only have demonstrative pronouns and zero as possible forms for 3rd person reference. What is at stake here is that in both languages the relevant forms of reference are ranked, from low to high accessibility in the order mentioned. Languages may thus differ in the granularity of accessibility distinctions that can be marked, but not in terms of workings of accessibility as a determinant of referential choice.

This also means that a certain form type such as a 3rd person pronoun is not as such associated with a particular degree of accessibility cross-linguistically; accessibility is in that sense relative and language-specific. An evaluation of AT and similar theories is likewise independent of both the propensity in a given language for either overt or zero reference, i.e. what Bickel (2003) calls 'referential density', and of the presence of fully systematic argument agreement in all persons for one or more syntactic functions in a given language. These two features have played a central role in research on reference production, but are not central to the question whether named theories make the right predictions about pronoun vs. zero use. Whether, say, a pronoun or zero be generally more frequent in a language, the former will — so the claim — *always* have a referent lower in accessibility than the latter, regardless of any other considerations.[3] Likewise the fact that a language may show agreement for a given argument function does not change anything in terms of the accessibility-driven ranking of a conominal pronoun vs. a zero argument, at least in principle. The only condition under which agreement might be relevant is where the paradigm shows a referentiality split. Most commonly, agreement is thus present in 1st and 2nd, but not 3rd persons, suggesting increased use of zero for the former due to complementarity. As will become clear below, this is not the pattern we are detecting in our data for which the presence of agreement is thus not relevant. Our finding, that more accessible referents show a greater likelihood of being realised as a pronoun does, however, have implications for the diachronic development of agreement, and we will return to this point briefly below.

2.2 The paradox of reduced forms

A number of case studies on high accessibility forms (pronouns and zero, here following Kibrik (2011), collectively "reduced forms") report findings that run

3. The gist of accessibility as a notion is that it subsumes any and all considerations that may be relevant to reference.

counter to the principle of maximum accessibility > minimal form. This is what we refer to here as the *paradox of reduced forms*. In what follows, we summarise some of the relevant literature, focussing on three categories that are considered to contribute to accessibility: animacy, person, and syntactic role.

Starting with animacy, Soares et al. (2020), building on extensive previous research, present the results of corpus-based and experimental studies on pronoun versus zero in Brazilian Portuguese (BP). They point to a consistent effect of animacy (human versus non-human), such that zero is more frequent with non-human referents than with human referents. This is unexpected on an Accessibility Theory approach, according to which zero should be preferred for the most accessible referents. Another study compared full versus (phonologically) reduced pronouns in two varieties of Dutch, employing an experimental methodology (Vogels et al. 2014). They find that "inanimate nouns are more likely to be referred to with the reduced forms than animate nouns." (Vogels et al. 2014: 117). Both studies thus independently confirm that it is the phonologically lighter member of the reduced forms that is preferred for referents lower in animacy, and hence in accessibility. As Vogels et al. (2014) note, this is not compatible with an accessibility-based approach, which assumes that "the more animate a conceptual representation of a referent is, the more attenuated the expression referring to that referent will be" (Vogels et al. 2014: 105). In a corpus-based study of Nepali discourse, Genetti and Crain (2003: 216) identify a strong bias towards "animate" referents among pronouns: more than 90% of pronouns in the corpus had animate referents. For "inanimates", then, the preferred reduced form is zero, leading them to formulate an "Animate pronoun constraint: Avoid pronominal mentions for inanimate referents".

Similarly, evidence from grammars suggests a reversal of the expected correlations. Meakins (2015) investigates the condition for clitic pronoun versus zero in Gurindji and Bilinarra, two languages of the Ngumpin subgroup of the Ngumpin-Yapa (Pama-Nyungan) family, Northern Territory, Australia. Among the factors affecting the realisation as a pronoun she points to person (first and second more likely than third, see below), but also animacy: third person non-human referents are more likely to have zero than human referents, which prefer overt pronominal form. Analogous findings are reported with regards to object realisation in a variationist study of text from the Oceanic language Vera'a (Schnell & Barth 2018): in narrative texts human objects are hardly ever realised as zero anaphors even where they occur in chains of coreferent objects. Conversely, non-human objects can hardly ever receive the form of a pronoun. Similarly, Movima, a language isolate from Bolivia, exhibits a tendency to realise inanimate P as zero (Haude & Allassonnière-Tang, to appear). These probabilistic constraints find their mirror image in hard constraints or grammatical rules in some languages, such as the

conditions on object drop reported by Skopeteas et al. (2024) for the Kwa languages Anyi and Baule. Ignoring a number of additional complications, as a general rule in both languages, inanimate objects cannot be realised as a pronoun and have to be dropped in given discourse contexts where animate ones by contrast have to be expressed as a pronoun; similar facts are reported for other Kwa languages. Examples (2a) and (2b) illustrate object drop in Anyi (examples from Skopeteas et al. (2024: 53–54), glossing slightly modified).

(2) a. Context: 'Do you have any news about the woman?'
*(blá dìɛ́) mɪ̀ mgbɪ̀dɛ̀-lì yí / *Ø*
woman TOP 1SG search-PFV 3SG
'(As for the woman,) I searched for her.'

b. Context: 'Do you have any news about the letter?'
*(klàtă dìɛ́) mɪ̀ mgbɪ̀dɛ̀-lí *yí / Ø*
letter TOP 1SG search-PFV 3SG
'(As for the letter,) I searched for it.'

The rule just outlined for object drop in Anyi will presumably impact the frequency of zeroes in discourse, leading to a greater number of zeroes for non-human objects than human objects, again countering the expected tendency. Thus we encounter in Portuguese, Dutch, Nepali, Vera'a, Movima, the Kwa languages of West Africa, and the Ngumpin-Yapa languages of North Australia a parallel trend: in grammar and in discourse, it is the lower-animacy referents that favour zero expression over an overt pronoun. While far from representative, the typological diversity and consistency of these findings constitute a significant challenge to a purely accessibility-based account of referential choice.

The second feature we consider is person, which is closely associated with animacy. First and second person forms refer by default to human entities, and should therefore be at least as accessible, if not more so, than human third person forms. Ariel (2000) goes further, claiming that first and second person forms are more accessible than third person forms, independent of animacy: "The general finding is then that first and second person referents are consistently highly accessible, but third person referents are only extremely highly accessible when they happen to be the continuing discourse topic(s)." Furthermore, "extremely high accessibility [...] remains constant for first and second person across all texts." (Ariel 2000: 221). On this view, then, first and second person forms are inherently more accessible than third person forms, even when their animacy values are identical (human).

In support of these claims, Ariel (2000) presents results from published corpus studies on conversational modern Hebrew. Table 1 summarises these studies, as cited in Ariel (2000). The figures indicate the proportion of zero form among

all reduced forms (pronoun+zero). Thus in Harel (1992), among all forms referring to first and second person referents, 31% are zero, while among the totality of reduced 3p forms (pronouns and zero), zero forms account for 17%. The difference between these two values is considered significant at an assumed significance threshold of 0.05.

Table 1. Proportions of zero reference across persons in four studies of Hebrew texts

Source	% zero among		Fisher's exact
	1&2pers	3pers	
Harel (1992), Tab. 1+2 in Ariel (2000)	31	17	$P = 0.007$
Levy (1995), Tab. 3 in Ariel (2000)	46	48	$P = 1.000$
Conversational data from Tab. 4+5 in Ariel (2000), source not provided	41	20	$P < 0.0001$
Ariel (1990), Tab. 6 in Ariel (2000)	31	22	$P = 0.0219$

Although these figures support an overall higher frequency of zero in the first and second person, the extensive literature on pro-drop in Spanish identifies a different tendency: first person subjects are less likely to be zero than other persons (Prada Pérez 2015). The overall figures from Genetti and Crain's study of Nepali discourse also speak against the accessibility account: 60% of first and second person reduced forms are zero, as opposed to 75% zero among third person reduced forms (Genetti & Crain 2003, Table 6). Schnell (2018) and Schnell and Barth (2020) report that zero subjects in the Oceanic language Vera'a are confined to the 3rd person, whereas 1st and 2nd persons are nearly always realised as pronouns. For earlier stages of Germanic, Rosenkvist (2009: 159) notes that across all varieties surveyed, zero is "by far the most frequent" for referential third person subjects, as opposed to first and second person. Direct comparison across different studies is difficult, as the Spanish, Germanic and Vera'a studies are generally restricted to subject drop, while Ariel's figures presumably include pronouns in all roles, as do Genetti and Crain's. Currently, then, it is not clear whether any cross-linguistically valid tendency for the impact of person can be identified. However, the findings from Spanish, Germanic, Nepali, and Vera'a already suffice to discount the universality of the findings based on the Hebrew data reported in Ariel (2000). We conclude that the accessibility-based account of person effects is not universally valid, and remains to be confirmed against more representative data.

The final factor we consider is syntactic role. There is widespread agreement that the subject role is associated with a higher degree of accessibility than the object role. For instance, Ariel (1990) considers the subject position to contextu-

ally raise salience, and hence the accessibility, of a discourse referent. This echoes work within Centering Theory (Grosz et al. 1995) and topic continuity (Chafe 1976; Givón 1983) which is in turn backed by numerous experimental studies on pronoun resolution (Stevenson et al. 2000; cf. Kehler & Rhode 2013; Kehler 2002). However, different studies interpret the role variable in different ways. In some (e.g. Bittner 2007), the crucial variable is the role of the antecedent of the token in question. Subject antecedents are predicted to trigger higher accessibility of anaphors than non-subject antecedents, and this is the variable that is manipulated in experimental settings. In other studies, the variable is the role of the token in question. This perspective is taken in the pro-drop literature, where a distinction is drawn between standard pro-drop, which affects subjects, and radical or discourse pro-drop, which additionally affects objects (Holmberg 2009; Neeleman & Szendröi 2007). There is a broad consensus that languages are more likely to license standard pro-drop (zero subjects) than zero object pronouns. From a discourse frequency perspective, this would translate to higher rates of zero among subject arguments than object arguments, a prediction that is largely confirmed in the Preferred Argument Structure literature (Du Bois 1987a; 2003; Haig & Schnell 2016 among others; cf. Chafe's (1994) 'Light subject constraint'). However, the strength of the effect and its interaction with person and animacy have not been systematically investigated. For example, the systematic association of non-human objects with zero, as illustrated in (2) above for Anyi, is a language-specific rule, and is thus not entirely accountable in terms of general cognitive principles. One problem is that in actual usage, role is entwined with other factors such as animacy and topicality, making it difficult to identify the specific contribution of role in determining form. Finally, there may also be an effect of same-role versus role-shift in determining referential choice in co-referential sequences ("parallelism effects", Arnold 2003: 237).

In sum, it is not straightforward to isolate the property 'being a direct object' from other factors in order to ascertain the impact of 'syntactic role' on the choice of reduced form. In the literature, different approaches and methodologies have been applied across various research traditions, making comparison across different studies difficult. In our case study below, we compare global frequencies of zeroes in spoken language corpora, contrasting tokens of transitive subject and direct object, but without considering the roles of antecedents.

2.3 Interim summary

Existing research on the correlation between high accessibility and reduced form yields conflicting results. With regard to animacy, several studies suggest that non-human referents are actually more likely to be associated with zero than with

overt pronouns. With regard to first and second person pronouns, Ariel (2000) cites findings from Hebrew that support the assumed correlation, but the pro-drop literature on Spanish and corpus-based work on Nepali and Vera'a point in the opposite direction. With regard to syntactic role, there is a tendency to associate subjects with zero realisations to a greater degree than objects, but the extent of cross-linguistic variation remains to be established. Furthermore, different researchers adopt different methodologies, some taking antecedent role as the dependent variable, while others consider the role of the token under consideration, and the basis for role comparison likewise differs (subjects versus only objects, or other non-subject roles?).

Overall, existing studies do not lend consistent support to the basic tenet of accessibility-based approaches, at least in the context of speakers' choices among reduced forms. This raises the possibility that it may not be meaningful to derive a single principle that drives referential choice across the entirety of the scale outlined in (1). Recent work by Haig et al. (2021) and Schiborr (2023) point to the existence of relatively consistent patterns determining the distribution of lexical NPs in discourse cross-linguistically, while the respective distribution of reduced forms (pronouns versus zero) varies dramatically across different languages. If these findings are on the right track, the determining factors behind the choice of reduced forms may not be entirely derivable from (presumably) universal aspects of processing and cognition, but are more susceptible to language-specific and historically contingent quirks of morphosyntax, such as gender and agreement systems, or the structure of pronominal paradigms. The rest of the paper is devoted to a cross-linguistic and corpus-based investigation that focusses specifically on the factors animacy, person, and role on the choice of form type among reduced forms.

3. Research questions and data sources

In what follows, we present results from an investigation of production data from 16 languages which have been annotated within a common framework. We address the following three general questions:

1. Are there cross-linguistically valid generalisations regarding the impact of animacy on the frequency of overt versus zero pronouns?
2. Are there cross-linguistically valid generalisations regarding the impact of person on the frequency of overt versus zero pronouns?
3. Are there cross-linguistically valid generalisations regarding the impact of syntactic role on the frequency of overt versus zero pronouns?

In the interest of brevity, we reduce each variable to a binary opposition:

Animacy:	human vs. non-human
Person:	first and second person (Speech Act Participant, SAP) vs. third person
Role:	transitive subject (A) vs. transitive object (P)

The data is from Multi-CAST (Haig & Schnell 2021), a multilingual corpus of spoken language texts. The texts are unscripted spoken monologues of minimally 1000 clauses of running text which have not been controlled for content; they are generally narratives, recounting personal events and oral history, or traditional tales; one corpus comprises Pear Film retellings (Chafe 1980) and hence represents a parallax corpus (cf. Barth & Evans 2017 for discussion). All data are fully accessible at https://fd-repo.uni-bamberg.de/records/jqgfp-vkj64. Currently, Multi-CAST contains texts from 20 diverse languages, with language-specific corpus sizes varying; an overview is given in Table 1.

The time-aligned transcriptions are accompanied by morphological glossing and additional annotation layers which primarily target discourse referents, providing information on grammatical relations (e.g. S, A and P), animacy, form type, and indexing coreference relationships across different mentions. Crucially, these annotations are also applied to referents with zero expression; see GRAID guidelines of morphosyntactic annotation (Haig & Schnell 2014) and RefIND guidelines on referent annotation (Schiborr, Schnell & Thiele 2018). For details on coding practices see Schnell et al. (2023). The annotations can be imported directly into R with the library *multicastR* for statistical analysis (Schiborr 2021).

All tokens of discourse referents in A and P function which have either zero or pronominal form were extracted from the data. Full lexical NPs are not included in the counts (but are listed in Table 3 as reference). This methodology replicates that of Ariel (2000) in her study of contemporary Hebrew (Table 1 above). The tokens thus identified were classified for person (SAP vs. third person) and animacy (SAP, and third person human vs. third person non-human). The dependent variable is form: zero vs pronominal. The sample of tokens is thus the totality of reduced forms identified in the corpus, and the relevant metric is the proportion of zero among the reduced forms in each category. An overview of the total data is provided in Table 3. Language-specific counts can be found in Appendix to this chapter.

Prior to analysis, we set a minimal threshold of 10 for the token numbers in each combination in each language. There are two combinations which yielded small numbers: 3p non-human referents in A function (seven languages were excluded) and SAP referents in P function (2 languages excluded). The overall paucity of tokens in these combinations is entirely expected, given what is known about the distribution of forms and functions in discourse (Haig & Schnell 2016;

Table 2. List of language subcorpora in Multi-CAST (cf. Schiborr 2025)

Language	Glottocode	Family	Clauses	Reference
Arta	arta1239	Austronesian	1,030	Kimoto 2019
Bora	bora1263	Boran	1,242	Seifart & Hong 2022
Chirag	chir1284	Nakh-Dagestanian	1,409	Ganenkov & Schiborr 2025
Cypriot Greek	cyor1249	Indo-European	1,070	Hadjidas & Vollmer 2015
English (Kent)	sout3282	Indo-European	5,649	Schiborr 2015
Jinghpaw	kach1280	Sino-Tibetan	1,278	Kurabe 2021
Kalamang	kara1499	Papuan	1,051	Visser 2021
Mandarin	mand1415	Sino-Tibetan	1,423	Vollmer 2020
Matukar Panau	matu1261	Austronesian	1,121	Barth, Davey, Matheas 2023
Nafsan	sout2856	Austronesian	1,012	Thieberger & Brickell 2019
Northern Kurdish	nort2641	Indo-European	1,841	Haig, Vollmer, Thiele 2019
Persian	tehr1242	Indo-European	1,418	Adibifar 2016
Sanzhi Dargwa	sanz1248	Nakh-Dagestanian	1,066	Forker & Schiborr 2019
Sumbawa	sumb1241	Austronesian	1,084	Shiohara 2022
Tabasaran	taba1259	Nakh-Dagestanian	1,382	Bogomolova, Ganenkov, Schiborr 2021
Teop	teop1238	Austronesian	1,303	Mosel & Schnell 2015
Tondano	tond1251	Austronesian	1,085	Brickell 2016
Tulil	taul1251	Papuan	1,264	Meng 2019
Uruangnirin	urua1244	Austronesian	1,079	Visser 2025
Vera'a	vera1241	Austronesian	3,608	Schnell 2015

Table 3. Absolute number of occurrences of reduced forms in A and P across 1st and 2nd person vs. 3rd person, and human vs. non-human, in contrast to (3rd person) lexical NPs

	Reduced forms				Lexical NPs	All discourse referents
	1+2p	3p hum	3p non-hum	Total reduced		
A	3647	4847	393	8887	1114	10001
P	604	1049	2298	3951	5714	9665
Total	**4251**	**5896**	**2691**	**12838**	**6828**	**19666**

Haig 2018; Haspelmath 2021; Aissen 2003; Comrie 1981). Figure 1 below is based on this version of the data.

4. Results

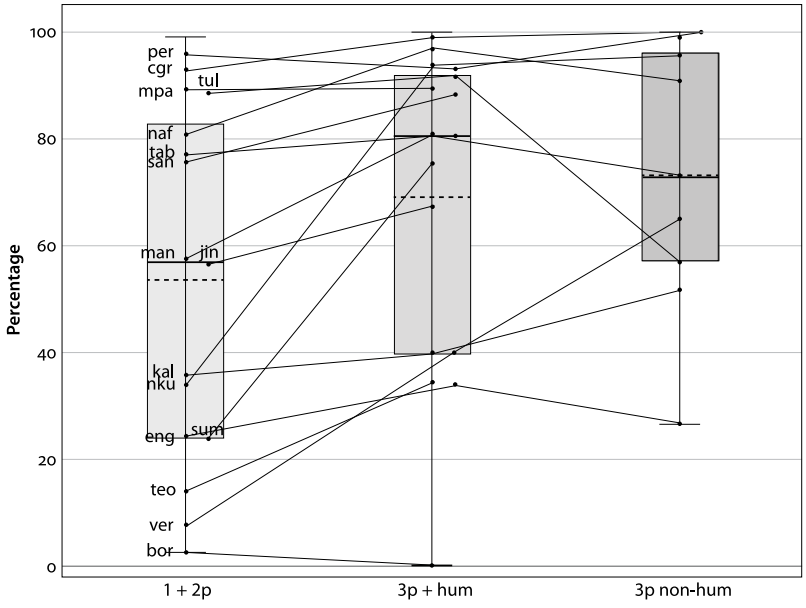
Figure 1 shows the respective frequencies of zero forms across three different person and animacy configurations, and split according to A versus P.

An initial comparison of the top and bottom panels indicates that overall, the mean value of zeroes is higher in A than P function, namely 66% (N=8853) of zero forms in A vs. 48% (N=3943) in P. This is entirely expected, on the assumption that the subject role is generally associated with topichood, widely considered a major driver of zero form. However, the effects of animacy and person are unexpected. In both A and P roles, we find the mean frequency of zero is lowest for SAP referents, and highest for third person non-human. The difference in the frequency of zeroes in SAP and third person human is obvious in P, but in A it also reaches significance.[4] The highest mean frequency of zero is with non-human third person, though the differences for A between human and non-human third person are not consistently skewed in the same direction, and difficult to assess due to low absolute figures for third person non-human A (seven data sets with n<10). In P, however, there is a consistent trend for higher rates of zero in non-human as opposed to human third person: in every language in the corpus, with the marginal exception of Matukar Panau, frequency of zeroes is highest for 3rd person non-human. In sum, we can confirm a general trend across both roles to avoid zero for SAP forms, and to prefer it for non-human third person. In A, the evaluation of human versus non-human third person is rendered difficult due to low absolute numbers and requires further confirmation. Our data provide no evidence for a correlation of high animacy with zero form, and substantial evidence for an inverse effect, such that low animacy favours zero form, most notably for P. These findings are not obviously compatible with accessibility or prominence-based approaches to referential choice.

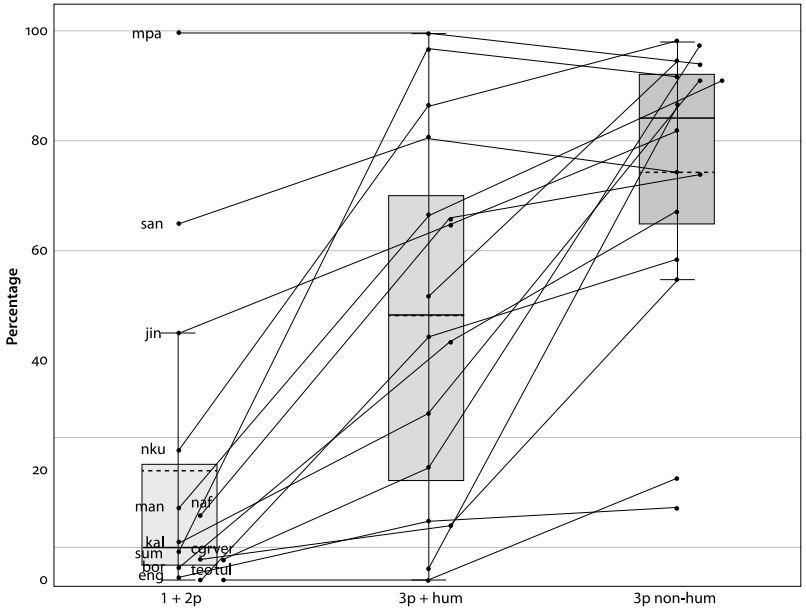
5. Discussion

Our findings counteract the expectations of Accessibility Theory, and as such dovetail with and expand on findings from previous studies questioning its pre-

4. A dependent t-test for SAP vs. 3p human in A function yields a t-statistic −4.859; p-value: 0.000231, confirming the higher mean frequency of zeroes for 3p human as significant.



A only, % zeroes



P only, % zeroes

Figure 1. Frequency of zeroes among reduced forms, according to person, animacy, and role across l6 corpora; dashed horizontal lines indicate mean, bold lines median values

dictive adequacy for high-accessibility referents. Both in regard to person and to humanness, we find that higher-ranked referents (SAPs and humans) are overall less likely to receive zero expression than lower-ranked referents (3rd persons and non-humans). Ignoring for a moment the role-related differences (see below), the overall results can be summed up in (3):

(3) Likelihood of zero as opposed to pronominal expression in spoken monologic discourse
3p non-hum > 3p hum > 1/2 person (SAP)

We assume that (3) implies a general principle for avoiding zero expression for human referents (SAP and third person human). Our data for forms of SAPs, which are highest on the animacy scale, are the most consistent in this regard,[5] so it is worth considering why they differ from those reported for contemporary Hebrew in Table 1 above, from Ariel (2000). There are several possible sources for the discrepancy. First: Ariel's data are mainly conversational, with absolute numbers of SAPs overwhelmingly *higher* than for third person, except in one corpus — which, crucially, did not show the trend expected by Ariel. Similar claims have been made for pro-drop in conversational English; Wagner (2018), summarising existing research on English, concludes that first person singular are the most frequently dropped forms, in line with Ariel's claims for Hebrew. However, most of the English studies are also based on conversational data, and the basis for the frequency counts vary across the studies considered. Our data, on the other hand, are predominantly monologic narratives, in which third person forms are overall far more frequent (Table 3), making up almost 80% of all discourse referents in the entire sample. Where SAPs do occur, they are essentially enacted, i.e. they figure in reported direct speech or thought that the narrator produces on behalf of the characters in the story. It is therefore possible that in conversational data, where the overall frequency of SAP pronouns is much higher, that they are more likely to be dropped, as suggested by the data from conversational English cited in Wagner (2018). This is a question for future research. With regard to the Hebrew findings, the results may also be related to presence of the rich agreement in Modern Hebrew in 1st and 2nd persons; this likewise requires further exploration.[6]

5. "Consistent" at least in the relative frequencies of zero found with different persons. We note that for five languages in our sample, SAP forms do show high frequencies (>80%) of zero in A function, but in these same languages, frequencies of zero are high across all person values. In other words, there are clearly some languages that just have an overall high frequency of zeroes in discourse, including for SAP. Interestingly, Mandarin is not among these languages, despite being considered a prototypical case of a language with radical or discourse pro-drop (Huang 2000); see Vollmer (2024) for recent discussion of Mandarin.

With regard to humanness, our findings echo those of Soares et al. (2020) and Vogels et al. (2014). While to our knowledge, no empirical evidence for the greater accessibility of human referents as opposed to nonhuman ones has been presented within Accessibility Theory, the theory does regard humans as inherently more salient. As salience is one of four key accessibility-lending features, humans are generally more accessible and should thus receive the highest accessibility marking, i.e. zero. It thus appears that considerations other than referent accessibility are at work when it comes to the realisation of human referents. One possibility is that salience bears its very own effects on referential choice: rather than simply raising accessibility, language users may be more likely to pair conceptual salience with phonologically heavier (and thus perceptually more salient) form, which would inhibit the use of zero. Another consideration is that in many languages, non-human referents do not receive overt marking for number when realised as a full NP. As discussed by Du Bois (1987b) this may carry over to the use of pronouns which are — across most languages — always inflected for number. It is thus in many languages not possible to use a singular pronoun with a non-singular reading, whereas zero anaphors are open to any number reading. As number is typically differentiated with human referents they may tend to be realised by non-singular pronouns where a multitude of humans is concerned, and this tendency towards pronoun use may carry over to singular instances as well. In a similar vein, Soares et al. (2020) consider the "Semantic Gender Hypothesis" in explaining the preference for zero among non-human referents in Portuguese. In essence, they suggest that the pronoun system of Portuguese obliges speakers to commit to a choice between masculine and feminine forms when referring to humans, cueing a corresponding semantic expectation for the recipient in pronoun resolution. In case of inanimate and non-specific referents, the choice of zero as anaphor leaves an inanimate reading as a natural (and perhaps preferred) option for the listener. They also note an effect of number, suggesting that pronouns can be analysed as realising feature combinations of [+/−plural] and [+/−animate], while zero is preferably associated with [−plural] and [−animate]. Pronouns are thus more informative than zero, but the nature of the categories expressed by pronouns varies cross-linguistically. While most have a number distinction, gender is by no means universal and even when present may have different values.

In particular for objects, the data from Multi-CAST exhibit a strong tendency in several languages to associate zero with a non-human referent. This is particularly notable for Persian, where only 2% of human objects are zero, while 87% of

6. A preliminary study of zero pronouns by Givón (2017: 86), based on contemporary written Hebrew fiction, does identify an effect of rich agreement on rates of pronoun deletion.

non-human objects are zero, or Vera'a, where the corresponding figures are 21% and 98% (neither language formally distinguishes humanness or gender in the pronominal paradigm). Given that gender systems tend to minimally implement the distinction between sex-differentiable human and non-sex-differentiable non-human (Corbett 1991), then it does not seem improbable that this distinction is replicated in discourse frequencies, where zero forms become loosely associated with non-human (or inanimate) discourse referents. These "soft" discourse tendencies have their counterpart in hard grammatical constraints on object drop, such as the one discussed above for Kwa languages (2a-2b), where non-human objects apparently cannot be pronominalised and instead must be zero, while human objects must take the form of a pronoun.

The second major finding of our study concerns the differences between A and P: the overriding principle associated with (3) manifests itself with greater magnitude in P than in A, in particular with respect to the humanness distinction in the third person. We surmise that this is related to differing baseline probabilities for referent types in different roles, which prime speakers' levels of predictability or expectancy for certain kinds of referents in certain roles. With regard to humanness, it has often been noted that As tend to be human simply because humans are overall much better candidates for fillers of agentive roles in transitive events (Haig, Schnell & Schiborr 2021; Haspelmath 2021). In our data, for example, over 95% of all A arguments are human (SAP or third person human). A non-human referent in A function is a comparatively rare event, and such deviations from the expected require additional morphological (and hence phonological) weight, say in the form of an affix that is not present in NPs with (expected) human reference. Explanations along these lines have been put forward in Haspelmath (2021), drawing on Zipf (1965), or the general principle in pragmatics that any more prolix form than is strictly necessary (e.g. by grammatical rules) will incur the inference of a less typical situation (Horn 1984). Such predictability / surprisal effects in a non-technical sense (Aissen 2003; Haspelmath 2021) would apply specifically to the realisation of non-human subjects in A function, due to the extremely low overall frequency of such items. That would constitute a conflict with (3), which states that low animacy (e.g. non-human) favours zero form over pronominal form. In P, on the other hand, the baseline frequency for non-human referents is much higher (more than 80%), so there is no conflict with the tendency in (3), and zero expression for non-human referents can apply unchecked.

The findings for SAP are more difficult to account for. It has been noted that overall frequencies of SAP pronouns are heavily dependent on content and text type (in-group informal conversations (high) versus traditional narratives (low), and so on). Some of our narrative texts contain very few SAP pronouns at all.

However, Haig (2018) shows that regardless of text type, the relative frequency of SAP pronouns in P function is consistently low (under 20%). This would provide some support for a surprisal effect in favouring pronominal over zero SAP in P function, which would be further amplified by the tendency in (3). But it is less clear how this generalises to A, where levels of zero for SAP pronouns are likewise lower than third person forms in all languages in our sample, with the sole exception of Persian, where zero SAP is marginally more frequent than zero third person. The low levels of zero for SAP pronouns are also surprising in view of the fact that many of the languages have overt agreement morphology for first and second person subjects (intransitive and transitive), which would presumably support zero as a general preference subjects in these languages. But in our data at least, these tendencies appear to be overridden by (3), providing further support for assuming that (3) is a very general tendency of discourse organisation.

In addition to surprisal, the deployment of pronouns for third person humans may also be due to their association with first and second person pronouns in a single morphological paradigm so that the consistent use of pronouns for SAPs primes language users to use what is essentially the same form type, a personal pronoun, for third person humans. This then also explains why we do not find a complete reversal in terms of markedness in the sense of Aissen (2003) for A: markedness reversal predicts least marked forms, i.e. zero in the current context, for expected, conceptually unmarked humans. But this is not the case as in our data, human As are oftentimes realised by a pronoun. This is likely due to the overriding principles guiding pronoun use, in particular the association of human referents with a person paradigm.

We also point to parallels with indexing systems (as a neutral terms for prosodically bound markers for replicating feature values of arguments; this would also include what is traditionally termed agreement morphology; see Corbett 2003; Haspelmath 2013, building on insights of Barlow 1992). The prevalence of zero indexing for the third person is typologically well-attested (but see Bickel et al. 2013 who question the validity of the assumed universal). More recently, Seržant and Moroz (2022) have re-formulated this tendency in terms of the low 'weight' of 3p index as opposed to first and second person. Thus cross-linguistically, indexing paradigms are organised in such a way that the heavier index is associated with the higher prominence, entirely in line with our frequency-based data on the realisation of pronouns and zero in discourse. The well-known diachronic link between free pronoun form use and grammaticalising person inflection points in the same direction: as 1st and 2nd persons pioneer the development (cf. Fuß 2005), this implies increased use of 1st and 2nd person pronouns to enter processes of morphologisation. Moreover, fully grammaticalised agreement systems are more likely to arise in A than P function (Haig 2018 and references therein). In the light of our

findings, this may be the result of the propensity for the A role to be filled with human participants which are in turn preferably realised as pronouns. This will naturally lead to a high base rate of pronoun use in A, and given the right structural conditions may lead to their morphologisation (Schnell under revision). A nicely illustrative example of such a putative development is reported in Himmelmann (1999) from Tagalog where in so-called 'undergoer voice' constructions, A arguments are nearly categorically realised as a pronominal clitic that attaches to the verb form; the argument is almost never dropped, even within coreference chains. The author does indeed suggest that in Tagalog undergoer voice, the A displays incipient agreement, and our findings suggest that the propensity of human pronouns in A may be a key ingredient in such a process across languages. Moreover, these considerations may also add to explaining to the discrepancies we found between A and P: as pronoun use may often be on the verge of becoming regularised in A this may lead to a sweeping effect, a generalisation that leads language users to extend the probabilistic rule of pronoun use to all A arguments, hence simplifying their production choices, and catalysing said morphologisation process (cf. Schnell under revision). Our findings from discourse thus converge with cross-linguistic research on agreement paradigms, suggesting a broader principle that human referents require overt linguistic forms, both in agreement morphology (if present in the language), and in the distribution of pronouns over zero in discourse.

Lastly, it is worth noting that the use of pronouns in contrast to zero has also been attributed to more idiosyncratic features, for instance the morphological shape of pronouns, as claimed relevant for the 'radical pro drop' properties of Mandarin Chinese by Neeleman and Szendröi (2007), or subject pronouns frequently fusing with auxiliaries in English (*I'm, we've, she's* etc.) which can impact on pro-drop, see Wagner (2018: 11; also Schnell & Barth 2020 for Vera'a). Whether these findings are overall verified, they do suggest that individual languages develop their own idiosyncratic rules of pronoun use. However, our findings nonetheless indicate the existence of a deeper cross-linguistic principle regarding the kind of reduced form that is appropriate for human referents — and it's not zero.

In sum, these findings give rise to an account of referential choice within high-accessibility markers where accessibility is not the only relevant consideration: as all form types involved will differentiate fairly minute differences in accessibility whose relevance for interlocutors may be questionable to begin with anyway, it seems likely that existing form differences are mobilised for other dimensions that are of communicative relevance. Thus we consider that the Accessibility Hierarchy in (1) requires revision, such that the accessibility factor deemed relevant for referential choice among full or lexical forms, or the choice between full and reduced forms, should not be generalised to account for choices among reduced

forms. More generally, the principle of 'higher accessibility implies phonologically lighter forms' does not apply among the reduced forms, at least insofar as high animacy is equated with higher accessibility.

6. Summary and outlook

Our point of departure has been the conflicting reports regarding the relationship between accessibility on the one hand, and the choice of zero versus pronominal forms on the other. While accessibility-based accounts assume that the same principles govern referential choices across both lexical and reduced forms, evident in the various versions of the Accessibility Hierarchy in (1) above, several case studies have failed to confirm these predictions, which we refer to as the paradox of reduced forms: when the choice is between pronouns and zero, the widely-assumed accessibility-lending factors do not consistently drive higher frequencies of zero.

Our own results from a more representative typological sample in fact suggest a reversal of the expected correlation among reduced forms: comparatively low-accessibility non-human referents prefer zero as opposed to pronouns, and high-accessibility first and second forms prefer pronouns to zero. This reversal holds both for transitive subject and for objects, but is much more pronounced for objects. We nevertheless emphasise that our data stem mostly from monological narrative texts, rather than conversational data; it is possible that conversational data may exhibit different regularities, particularly with regard to the feature of person.

In addition, our findings suggest that differentiating between different syntactic contexts, in this case the syntactic function, is vital to obtain a realistic picture of referential choice. More generally, this observation points in an important direction for future studies into referential choice: referential choice is not only a probabilistic but also a multi-factorial matter, and rather than relying too heavily on an abstract aggregate concept like accessibility, it would seem more promising to consider relevant functional and structural constellations of discourse referents directly, also as these ultimately act as cues for speakers to draw on during their decision making process. A relevant avenue for this research is the modelling of referential choice with computational methods, as in Same (2024) or Linders and Schnell (under revision). Whether distinct weighted factors will lend themselves to the postulation of more generalisable principle akin to accessibility will thus be a matter of empirical enquiry, rather than an a priori assumption.

References

Aissen, Judith. 2003. "Differential Object Marking. Iconicity vs. Economy." *Natural Language and Linguistic Theory* 21 (3): 435–483.

Ariel, Mira. 1988. "Referring and Accessibility." *Journal of Linguistics* 24 (1): 65–87.

Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents*. London: Routledge.

Ariel, Mira. 2000. "The Development of Person Agreement Markers: From Pronouns to Higher Accessibility Markers." In *Usage-based Models of Language*, ed. by Suzanne Kemmer, and Michael Barlow, 197–261. Stanford: CSLI Publications.

Arnold, Jennifer E. 1998. Reference Form and Discourse Patterns. Doctoral dissertation. Stanford, CA: Stanford University.

Arnold, Jennifer E. 2003. "Multiple Constraints on Reference Form." In *Preferred Argument Structure: Grammar as Architecture for Function*, ed. by John W. DuBois, Lorraine Kumpf, and William Ashby, 225 — 246. Amsterdam: John Benjamins.

Arnold, Jennifer E. 2008. "Reference Production. Production-Internal and Addressee-Oriented Processes." *Language and Cognitive Processes* 23 (4): 495–527.

Arnold, Jennifer E. 2010. "How Speakers Refer. The Role of Accessibility." *Languages and Linguistics Compass* 4 (4): 187–203.

Arnold, Jennifer E., Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. "Heaviness vs. Newness. The Effect of Structural Complexity and Discourse Status on Constituent Ordering." *Language* 76 (1): 28–55.

Adibifar, Shirin. 2016. "Multi-CAST Persian." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. Version 2507", ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg. ()

Barlow, Michael. 1992. *A Situated Theory of Agreement*. London: Routledge.

Barth, Danielle, and Nicholas Evans. 2017. *The Social Cognition Parallax Corpus (SCOPIC)*. Language Documentation & Conservation Special Publication No. 12. http://hdl.handle.net/10125/24739

Barth, Danielle, Kira Davey, and Maria Matheas. 2023. "Multi-CAST Matukar-Panau." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Bickel, Balthasar. 2003. "Referential Density in Discourse and Syntactic Typology." *Language* 79 (4): 708–736.

Bickel, Balthasar, Giorgio Iemmolo, Taras Zakharko, and Alena Witzlack-Makarevich. 2013. "Patterns of Alignment in Verb Agreement." In *Languages Across Boundaries. Studies in Memory of Anna Siewierska*, ed. by Dik Bakker, and Martin Haspelmath, 15–36. Berlin: Mouton de Gruyter.

Bisang, Walter. 2014. "Overt and Hidden Complexity. Two Types of Complexity and Their Implications." *Poznań Studies in Contemporary Linguistics* 50 (2): 127–143.

Bisang, Walter. 2015. "Hidden Complexity. The Neglected Side of Complexity and its Implications." *Linguistics Vanguard* 1 (1): 177–187.

Bittner, Dagmar. 2007. "Influence of Animacy and Grammatical Role on Production and Comprehension of Intersentential Pronouns in German L1-Acquisition." *ZAS Papers in Linguistics* 48: 103–138.

Bogomolova, Natasha, Dmitry Ganenkov, and Nils N. Schibor. 2021. "Multi-CAST Tabasaran." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Boswijk, Vincent, and Matt Cobler. 2020. "What is Salience?" *Open Linguistics* 6: 713–722.

Vogels, Jorrig, Alfons Maes, and Emil Krahmer. 2014. "Choosing Referring Expressions in Belgian and Netherlandic Dutch: Effects of Animacy."*Lingua* 145: 104–121.

Brickell, Timothy. 2016. Multi-CAST Tondano. In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts* ed. By Geoffrey Haig and Stefan Schnell. Bamberg: University of Bamberg.

Chafe, Wallace. 1976. "Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View." In *Subject and Topic*, ed. by Charles N. Li, 27–55. New York: Academic Press.

Chafe, Wallace. 1980. *The Pear Stories. Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex.

Chafe, Wallace. 1987. "Cognitive Constraints on Information Flow." In *Coherence and Grounding in Discourse*, ed. by Russel S. Tomlin, 21–51. Amsterdam: John Benjamins.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time. The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago, IL: The University of Chicago Press.

Clancy, Patricia M. 1980. "Referential Choice in English and Japanese Narrative Discourse." In *The Pear Stories. Cognitive, Cultural, and Linguistic Aspects of Narrative Production*, ed. by Wallace Chafe, 127–202. Norwood, NJ: Ablex.

Clark, Herbert H., and Gregory L. Murphy. 1982. "Audience Design in Meaning and Reference." In *Language and Comprehension*, ed. by Jean-François Le Ny, and Walter Kintsch, 287–299. Amsterdam, New York, Oxford: North-Holland.

Comrie, Bernard. 1981. *Language Universals and Linguistic Typology*. Oxford: Blackwell.

Corbett, Greville C. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville C. 2003. "Agreement: Terms and Boundaries." In *The Role of Agreement in Natural Language: TLS 5 Proceedings*, ed. by W.E. Griffin, 109–122.

Du Bois, John W. 1987a. "The Discourse Basis of Ergativity." *Language* 63 (4): 805–855.

Du Bois, John W. 1987b. "Absolutive Zero. Paradigm Adaptivity in Sacapultec Maya." *Lingua* 71 (2): 203 — 222.

Du Bois, John W. 2003. "Argument Structure. Grammar in Use." In *Preferred Argument Structure. Grammar as Architecture for Function*, ed. by John W. Du Bois, Lorraine Kumpff, and William J. Ashby, 11–60. Amsterdam: John Benjamins.

Duranti, Alessandro. 1994. *From Grammar to Politics: Linguistic Anthropology in a Western Samoan Village*. Oakland: University of California Press.

Ganenkov, Dmitry, and Nils N. Schiborr. 2025. "Multi-CAST Chirag." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig and, Stefan Schnell. Bamberg: University of Bamberg.

Genetti, Carol, and Laura D. Crain. 2003. "Beyond Preferred Argument Structure: Sentences, Pronouns and Given Referents in Nepali." In *Preferred Argument Structure: Grammar as Architecture for Function*, ed. by John W. Du Bois, Lorraine E. Kumpf, and William J. Ashby, 197–223. Amsterdam: John Benjamins.

Gibson, Edward, Richard Futrell, Steven T. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. "How Efficiency Shapes Human Language." *Trends in Cognitive Science* 23 (12): 389–407.

Givón, Talmy. 1983. "An Introduction." In *Topic Continuity in Discourse*, ed. by Talmy Givón, 1–42. Amsterdam: John Benjamins.

Givón, Talmy. 2017. *The Story of Zero*. Amsterdam/Philadelphia: John Benjamins.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. "Centering. A Framework for Modelling the Local Coherence of Discourse." *Computational Linguistics* 21 (2): 203–225.

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 1993. "Cognitive Status and the Form of Referring Expressions in Discourse." *Language* 69 (2): 274–307.

Hadjidas, Harris, and Maria Vollmer. 2015. "Multi-CAST Cypriot Greek." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Haig, Geoffrey. 2018. "The Grammaticalization of Object Pronouns: Why Differential Object Indexing is an Attractor State." *Linguistics* 56 (4): 781–818.

Haig, Geoffrey, and Stefan Schnell. 2014. *Annotations Using GRAID: Grammatical Relations and Animacy in Discourse*. Bamberg: University of Bamberg.

Haig, Geoffrey, and Stefan Schnell. 2016. "The Discourse Basis of Ergativity Revisited." *Language* 92 (3): 591–618.

Haig, Geoffrey, and Stefan Schnell. 2021. *Multi-CAST. Multilingual Corpus of Annotated Spoken Texts*. Version 2507. Bamberg: University of Bamberg.

Haig, Geoffrey, Stefan Schnell, and Nils N. Schiborr. 2021. "Universals of Reference in Discourse and Grammar: Evidence from the Multi-CAST Collection of Spoken Corpora." In *Doing Corpus-Based Typology with Spoken Language Data: State of the Art*, ed. by Geoffrey Haig, Stefan Schnell, and Frank Seifart, 141–177. Honolulu, HI: University of Hawai'i Press. http://hdl.handle.net/10125/74660

Haig, Geoffrey, Maria Vollmer, and Hanna Thiele. 2019. "Multi-CAST Northern Kurdish." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Harel, Zvi. 1992. "Yossi, Believe Me as You Never Believed Me Before." *The Haaretz Supplement*, 13 Nov 1992, 20–25.

Haspelmath, Martin. 2013. "Argument Indexing. A Conceptual Framework for the Syntactic Status of Bound Person Forms." In *Languages across Boundaries. Studies in Memory of Anna Siewierska*, ed. by Dik Bakker, and Martin Haspelmath, 197–226. Berlin: Mouton de Gruyter.

Haspelmath, Martin. 2021. "Explaining Grammatical Coding Asymmetries. Form-Frequency Correspondences and Predictability." *Journal of Linguistics* 57 (3): 605–633.

Haude, Katharina, Marc Allassonière-Tang. to Appear. "Lexical, Pronominal and Zero Argument Encoding in Movima." In *Topicality and the Shaping of Grammar*, ed. by Claudine Chamoreau, and Enrique Palancer. Amsterdam: John Benjamins.

Himmelmann, Nikolaus P. 1997. *Deiktikon, Artikel, Nominalphrase. Zur Emergenz syntaktischer Struktur*. Tübingen: Niemeyer.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 1999. "The Lack of Zero Anaphora and Incipient Person Marking in Tagalog." *Oceanic Linguistics* 38 (2): 231–269.

Himmelmann, Nikolaus P. 2012. "Linguistic Data Types and the Interface between Language Documentation and Description." *Language Documentation & Conservation* 6: 187–207. http://hdl.handle.net/10125/4503

Himmelmann, Nikolaus P. 2014. "Asymmetries in the Prosodic Parsing of Function Words: Another Look at the Suffixing Preference." *Language* 90 (4): 927–960.

Himmelmann, Nikolaus P. 2022. "Prosodic Phrasing and the Emergence of Syntactic Structure." *Linguistics* 60 (3): 715–743.

Himmelmann, Nikolaus P., and Beatrice Primus. 2015. "Prominence Beyond Prosody — A First Approximation." In *Prominences in Linguistics. Proceedings of the pS-prominenceS International Conference*, ed. by Amedeo De Dominicis, 38–58. Viterbo: DISUCOM Press.

Holmberg, Anders. 2009. "Null Subject Parameters." In *Parametric Variation*, ed. by Theresa Bieberauer, Anders Holmberg, Ian Roberts, and Michelle Sheehan, 88–124. Cambridge: Cambridge University Press.

Horn, Lawrence. 1984. "Toward a New Taxonomy for Pragmatic Inference: Q-Based and R-Based Implicature." In *Meaning, Form, and Use in Context: Linguistic Applications*, ed. by D. Schiffrin, 11–42. Georgetown University Press.

Huang, Yan. 2000. *Anaphora*. Oxford: Oxford University Press.

Karttunen, Lauri. 1976. "Discourse Referents." In *Syntax and Semantics 7. Notes from the Linguistic Underground*, ed. by James D. McCawley, 363–385. New York: Academic Press.

Kehler, Andrew. 2002. Coherence, Reference, and the Theory of Grammar. PhD dissertation. Stanford, CA: CSLI Publications.

Kehler, Andrew. 2006. "Discourse Coherence." In *Handbook of Pragmatics*, ed. by Laurence Horn and Gregory Ward, 241–265. Malden, MA: Blackwell.

Kehler, Andrew, and Hannah Rhode. 2013. "A Probabilistic Reconciliation of Coherence-Driven and Centering-Driven Theories of Pronoun Interpretation." *Theoretical Linguistics* 39 (1–2): 1–37.

Kehler, Andrew, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman. 2008. "Coherence and Coreference Revisited." *Journal of Semantics* 25 (1): 1–44.

Kibrik, Andrej A. 2011. *Reference in Discourse*. Oxford: Oxford University Press.

Kimoto, Yukinori. 2019. "Multi-CAST Arta." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Kurabe, Keita. 2021. "Multi-CAST Jinghpaw." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Levy, Gideon. 1995. "That I Should Have to Defend Arafat?" *The Ha'aretz Supplement*. February 17, 1995, 14–16.

Levy, Roger, and T. Florian Jäger. 2007. "Speakers Optimize Information Density through Syntactic Reduction." In *Adavances in Neural Information Processing Systems 19. Proceedings of the 2006 Conference*, ed. by Bernard Schölkopf, John Platt, and Thomas Hoffmann, 849–856. Cambridge, MA: MIT Press.

Linders, Guido M., and Stefan Schnell. Under review. "Functional factors predict referential choice similarly across languages: A cross-linguistic computational analysis." Preprint at *PsyArXiv*.

Meakins, Felicity. 2015. "Not Obligatory. Bound Pronoun Variation in Gurindji and Bilinarra." *Asia-Pacific Language Variation* 1 (2): 128–162.

Forker, Diana, and Nils N. Schiborr. 2019. "Multi-CAST Sanzhi Dargwa." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. Version 2507", ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg. ()

Meng, Chenxi. 2019. "Multi-CAST Tulil." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Mosel, Ulrike, and Stefan Schnell. 2015. "Multi-CAST Teop." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig and Stefan Schnell. Bamberg: University of Bamberg.

Neeleman, Ad, and Kriszta Szendröi. 2007. "Radical Pro Drop and the Morphology of Pronouns." *Linguistic Inquiry* 38 (4): 671–714.

Patterson, Clare, and Petra B. Schumacher. 2020. "The Timing of Prominence Information during the Resolution of German Personal and Demonstrative Pronouns." *Dialogue & Discourse* 11 (1): 1–39.

Prada Pérez, Ana De. 2015. "First Person Singular Subject Pronoun Expression in Spanish in Contact with Catalan." In *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*, ed. by Ana M. Carvalho, Rafael Orozco, and Naomi L. Shin, 121–142. Washington, DC: Georgetown University Press.

Rosenkvist, Henrik. 2009. "Null Referential Subjects in Germanic Languages — An Overview." *Working Papers in Scandinavian Syntax* 84: 151–180.

Rosenkvist, Henrik. 2018. "Null Subject and Distinct Agreement in Modern Germanic." In *Null Subjects in Generative Grammar. A Synchronic and Diachronic Perspective*, ed. by Federica Cognola and Jan Casalincchio, 285–306. Oxford: Oxford University Press.

Same, Fahime. 2024. *Referring Expression Generation in Context. Combining Linguistic and Computational Approaches*. Berlin: Language Science Press.

Schiborr, Nils. N. 2015. "Multi-CAST English." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Schiborr, Nils N. 2021. "MulticastR. A Companion to the Multi-CAST Collection. R package version 2.0.0." In *Multi-CAST. Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. https://cran.r-project.org/package=multicastR

Schiborr, Nils N. 2023. *Lexical Anaphora. A Corpus-Based Typological Study of Referential Choice*. Bamberg: University of Bamberg Press.

Schiborr, Nils N. 2025. "Multi-CAST Collection Overview." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Schiborr, Nils N., Stefan Schnell, and Hannah Thiele. 2018. *RefIND — Referent Indexing in Natural-language Discourse. Annotation guidelines (v1.1)*. Bamberg: University of Bamberg.

Schnell, Stefan. 2015. "Multi-CAST Vera'a." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Schnell, Stefan. 2018. "Whence Subject-Verb Agreement? Investigating the Role of Topicality, Accessibility, and Frequency in Vera'a Texts." *Linguistics* 56 (4): 735–780.

Schnell, Stefan. under revision. "The Person Indexing Cycle. A Source-Oriented View from Corpus-Based Typology." *Language and Linguistics Compass*.

Schnell, Stefan, and Danielle Barth. 2018. "Discourse Motivations for Pronominal and Zero Objects Across Genres in Vera'a." *Language Variation and Change* 30 (1): 51–81.

Schnell, Stefan, and Danielle Barth. 2020. "Expression of Anaphoric Subjects in Vera'a. Functional and Structural Factors in the Choice between Pronoun and Zero." *Language Variation and Change* 32 (3): 267–291.

Schnell, Stefan, Geoffrey Haig, Nils N. Schiborr, and Maria Vollmer. 2023. "Are Referent Introductions Sensitive to Forward Planning in Discourse?" In *Discourse Phenomena in Typological Perspective*, ed. by Alessandra Barotto, and Simone Mattiola, 231–268. Amsterdam/Philadephia: John Benjamins.

Seifart, Frank, and Tai Hong. 2022. "Multi-CAST Bora." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Seržant, Ilja, and George Moroz. 2022. "Universal Attractors in Language Evolution Provide Evidence for the Kinds of Efficiency Pressures Involved." *Humanities and Social Sciences Communications* 9 (58).

Shiohara, Asako. 2022. "Multi-CAST Sumbawa." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Skopeteas, Stavros, Firmin Ahoua, Marie Laure Koffi Bla Adou, and Beatrice Koffi Mambo. 2024. "Resuming Topics and Foci: Anyi, Baule and Microvariation in Kwa Languages." *Journal of African Languages and Linguistics* 45 (1): 45–84.

Soares, Eduardo Correa, Philip Harold Miller, and Barbara Hemforth. 2020. "The Effect of Semantic and Discourse Features on the Use of Null and Overt Subjects — A Quantitative Study of Third Person Subjects in Brazilian Portuguese." *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 36 (1): 1–38.

Stevenson, Rosemary J., Alistair Knott, Jon Oberlander, and Sharon McDonald. 2000. "Interpreting Pronouns and Connectives. Interactions among Focussing, Thematic Roles, and Coherence Relations." *Language and Cognitive Processes* 15 (3): 225–262.

Stoll, Sabine, and Balthasar Bickel. 2009. "How Deep are Differences in Referential Density?" In *Crosslinguistic Approaches to the Psychology of Language. Research in the Tradition of Dan Isaac Slobin*, ed. by Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura, and Şeyda Özçalışkan, 543–555. London: Psychology Press.

Thieberger, Nicholas, and Timothy Brickell. 2019. "Multi-CAST Nafsan." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Tily, Harry, and Steven Piantadosi. 2007. "Refer Efficiently. Use Less Informative Expressions for More Predictable Meanings." In *Proceedings of the Workshop on the Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference (PRE-CogSci 2009)*, ed. by Kees Van Deemter, Albert Gatt, Roger P. G. van Gompel, and Emiel Krahmer. Tilburg: University of Tilburg.

Torres Cacoullos, Rena, and Catherine E. Travis. 2019. "Variationist Typology. Shared Probabilistic Constraints across (Non-)Null Subject Languages." *Linguistics* 57 (3): 653–692.

Visser, Eline. 2021. "Multi-CAST Kalamang." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Visser, Eline. 2025. "Multi-CAST Uruangnirin." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Vollmer, Maria. 2020. "Multi-CAST Mandarin." In *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*, ed. by Geoffrey Haig, and Stefan Schnell. Bamberg: University of Bamberg.

Vollmer, Maria. 2024. "Comparing Zero and Referential Choice in Eight Languages with a Focus on Mandarin Chinese." *Studies in Language* 48 (2): 351–389.

Von Heusinger, Klaus, and Petra B. Schumacher. 2019. "Discourse Prominence: Definition and Application." *Journal of Pragmatics* 154: 117–127.

Wagner, Susanne. 2018. "Never Saw One — First-Person Null Subjects in Spoken English." *English Language and Linguistics* 22 (1): 1–34.

Zipf, George K. 1965. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.

Absolute figures for Figure 1. Reduced (non-lexical) forms only, split for A and P, values <10 removed

A only	bor	cgr	eng	jin	kal	man	mpa	naf	nku	per	san	sum	tab	teo	tul	ver
total SAP	**75**	**250**	**1415**	**151**	**112**	**124**	**232**	**111**	**308**	**24**	**124**	**101**	**172**	**104**	**166**	**178**
zero SAP	3	232	360	86	41	72	207	90	108	23	94	25	133	16	147	16
%zero SAP	4	92,2	25,5	57,7	36,6	58,8	89,9	81,1	35,5	95,5	75,5	24,4	77,7	15,5	88,8	9,0
total 3p+hum	231	191	720	258	164	253	260	247	302	461	156	172	354	253	170	655
zero 3p+hum	4	189	252	175	67	205	233	239	283	429	138	130	286	90	156	267
%zero 3p+hum	1,7	99,9	35,5	67,7	40,0	81,1	89,9	96,6	93,3	93,3	88,8	75,5	80,0	35,5	91,1	40,0
total 3p non-hum	(<10)	23	177	(<10)	21	15	(<10)	33	23	15	(<10)	(<10)	(<10)	(<10)	26	26
zero 3p non-hum		23	49		11	11		30	22	15					15	17
%zero 3p non-hum		100	27,7		52,2	73,3		90,0	95,5	100					57,7	65,5

P only	bor	cgr	eng	jin	kal	man	mpa	naf	nku	per	san	sum	tab	teo	tul	ver
total SAP	44	26	158	40	29	15	25	25	67	(<10)	20	20	(<10)	10	38	79
zero SAP	1	1	1	18	2	2	25	3	16		13	1		0	0	3
%zero SAP	2,3	3,8	0,6	45,0	6,9	13,3	100,0	12,0	23,9		65	5		0	0	3,8
total 3p+hum	85	60	93	68	49	42	65	53	82	49	26	32	27	99	53	166
zero 3p+hum	37	6	10	44	15	28	65	35	71	1	21	31	14	44	0	34
%zero 3p+hum	43,5	10,0	10,8	64,7	30,6	66,7	100,0	66,0	86,6	2,0	80,8	96,9	51,9	44,4	0,0	20,5
total 3p non-hum	55	131	674	145	116	79	156	97	137	105	63	75	98	92	117	158
zero 3p non-hum	37	72	90	119	106	72	147	72	135	91	47	69	93	54	22	154
%zero 3p non-hum	67,3	55,0	13,4	82,1	91,4	91,1	94,2	74,2	98,5	86,7	74,6	92,0	94,9	58,7	18,8	97,5

Input data (percentages zero) for Figure 1

A	bor	cgr	eng	jin	kal	man	mpa	naf	nku	per	san	sum	tab	teo	tul	ver
SAP	4,0	92,8	25,4	57,0	36,6	58,1	89,2	81,1	35,1	95,8	75,8	24,8	77,3	15,4	88,6	9,0
3p+hum	1,7	99,0	35,0	67,8	40,9	81,0	89,6	96,8	93,7	93,1	88,5	75,6	80,8	35,6	91,8	40,8
3p non-hum	–	100,0	27,7	–	52,4	73,3	–	90,9	95,7	100,0	–	–	–	–	57,7	65,4
P	**bor**	**cgr**	**eng**	**jin**	**kal**	**man**	**mpa**	**naf**	**nku**	**per**	**san**	**sum**	**tab**	**teo**	**tul**	**ver**
SAP	2,3	3,8	0,6	45,0	6,9	13,3	100,0	12,0	23,9	–	65,0	5,0	–	0,0	0,0	3,8
3p+hum	43,5	10,0	10,8	64,7	30,6	66,7	100,0	66,0	86,6	2,0	80,8	96,9	51,9	44,4	0,0	20,5
3p non-hum	67,3	55,0	13,4	82,1	91,4	91,1	94,2	74,2	98,5	86,7	74,6	92,0	94,9	58,7	18,8	97,5

CHAPTER 20

A global view on the *Ausrichtungsaffix* and the deep differences between nominalization-derived versus applicative-derived diathesis in Austronesian

Daniel Kaufman
Queens College, City University of New York

Austronesian voice has been claimed to originate in an ancient reanalysis of participant nominalization as matrix clause predicates. A survey of participant nominalization as a relativizing strategy reveals that Austronesian voice has many analogues throughout the world that show certain typological hallmarks such as apparently null-headed relative clauses, genitive marked agents and a lack of dedicated relativizers and copulas. Many Austronesian languages maintain remnants of the original verbal forms alongside the nominalized forms that took over main clause functions in most languages. New observations about the case assignment properties of these forms in several Philippine languages strengthen the argument that they remain as a distinct verbal syntactic category in modern languages, in comparison to the participle-like forms stemming from nominalization.

Keywords: Austronesian voice, nominalization, relative clauses, grammaticalization, diathesis

1. Introduction

In this contribution, I argue for the *Ausrichtungsaffix* ('orientation affix') of Himmelmann (1987) as a distinct typological category that comes about through the reinterpretation of participant nominalization in a diathesis function, as originally proposed by Starosta et al. (1982/2009) to account for the nature of what has come to be known as "Philippine-type voice". In the time since these pioneering proposals, a large amount of cross-linguistic data has come to light which shows

https://doi.org/10.1075/slcs.240.20kau
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

that diathesis alternations rooted in participant nominalization leave a distinct typological fingerprint. In other words, the Austronesian family is far from alone in possessing a "Philippine type voice" built on participant nominalizations (cf. Sasse 2009).

The second point I discuss here may at first seem like a parochial concern of Austronesianists but turns out to have a strong impact on the notion of orientation morphology as a coherent type. Wolff (1973) reconstructs two seemingly unrelated sets of verbal morphology for Proto-Austronesian which he terms the dependent and independent paradigms, later recast by Ross (2009) as the first generation and second generation verb forms. The dependent/first generation paradigm, according to Starosta et al. (1982/2009) and Ross (2009), represent the original verbal morphology of Proto-Austronesian while the independent/second-generation forms represent a set of nominalizations or participles that were originally used in relative clauses but which later came to be reinterpreted as main clause predicates. In many Austronesian languages, both paradigms continue to be employed in distinct syntactic contexts although only the Formosan language Puyuma, according to Ross (2009), approximates the original distribution of labor, with first generation forms generally restricted to main clauses and second generation forms generally restricted to relative clauses and nominalizations.

In earlier work, I raised the possibility that remnants of first generation forms, even in languages like Tagalog, where they are highly restricted, still behave more like canonical verbs in comparison to their second generation counterparts (Kaufman 2009a). In response, Himmelmann (2009) notes the need to further flesh out these poorly understood facts:

> For other languages, including possibly southern Tagalog dialects, the analysis for nominal predicates in equational structures needs to be complemented with an analysis of a more verbal predicate type (as it stands, it is unclear how K[aufman] would analyse *i*-suffixed predicates in southern Tagalog and the more clearly verbal dependent forms attested in most other Philippine-type languages […]). Such an analysis would have to be able to account for the fact that, at least in some languages and constructions, the marking of the arguments of the more verb-like predicates is identical with the marking of nominal predicates (i.e. in particular the genitive-marking of actors of undergoer voice verbs). But, at the same time, it has to allow verbal heads also to occur with cross-linguistically more canonically marked arguments (e.g. as direct ('accusative') complements or, in undergoer voice constructions, as oblique agents), as such marking is widespread in the symmetrical voice languages of Indonesia.

I attempt to answer some of these questions here, specifically in relation to case marking, by looking more closely at the syntactic behavior of dependent forms in grammatically conservative Philippine languages. In doing so, I uncover a new

generalization: Whereas the presence of an independent paradigm always implicates the presence of a nominative argument (whether overt or elided), there is no such requirement on dependent forms. In other words, the dependent paradigm, across a range of constructions, can show diathesis alternations in the absence of a nominative pivot. This in turn suggests that nominative case assignment is dependent on finiteness in the dependent paradigm but simply reliant on an equational clause in the independent paradigm. This can be interpreted in several ways but in combination with other facts discussed below, I will argue that the dependent paradigm, unlike the independent paradigm, is crucially linked to higher functional categories associated with verbal inflection cross-linguistically.

In arguing for these deeper differences between the two Austronesian paradigms, I also argue implicitly for Starosta et al.'s (1982/2009) original claim that the independent paradigm originates in a set of participant nominalizations. While a link between relative clause predicates and nominalization has been noted in the syntax literature for individual Austronesian languages (among others), it is just as often dismissed as relic homophony without a synchronic basis (e.g., Chung 1994: 9). My arguments for a synchronic categorial difference between the two paradigms also stand in contrast to a more recent idea that these paradigms simply represent a difference in Mood, which Chen (2017: 241) proposes as the historical basis for this split:

> The Nominalization-into-verb (Nom-into-V) hypothesis was a misinterpretation of the fact that relative clauses are seldom formed in non-indicative moods, and thus rarely contain non-indicative voice morphology. Since the voice affixes in relative clauses are conventionally called nominalizers, previously works have incorrectly assumed a diachronic relation between Philippine-type indicative voice affixes and 'nominalizers'.

In the following section, I show that thematic nominalization is an extremely widespread relativization strategy, far more so than the use of a relative pronoun/wh-expression coreferent with the gap in a finite clause, which has already been shown to be quite restricted in its geographic scope (cf. Comrie 1981; Comrie & Estrada-Fernández 2012; Keenan 1985). Matisoff (1972), Watters (1978) and Mazaudon (1978) were among the first to note the prominence of nominalization in Tibeto-Burman relative structures and the same has now been shown for Andean languages (Lefebvre & Muysken 1988; Gijn 2024), Uto-Aztecan languages (Álvarez González 2012), Carib languages (Gildea 1998), among many others, forming a large body of evidence that was not available to Starosta et al. (1982/2009) at the time. The pathway from thematic nominalization to relativization is clear; what begins as an appositive structure, e.g. '[the woman], [the driver of the car]' is reinterpreted as a subordination, e.g. '[the woman [driver of the car]]', where a nominalizer like English *-er* is now specialized to mark predicates from

which the subject has been relativized.[1] A further step, in which these nominalized forms, originally relegated to modification but now employed for simple main clause predications, is also well attested and belongs to a wider family of phenomena termed "insubordination" (Evans & Watanabe 2016).

2. The two diathesis paradigms of early Austronesian

Building on Wolff (1973), Ross (2002) reconstructs the core proto-Austronesian voice paradigm as shown in Table 1. While this is greatly elaborated upon in Ross (2009), the facts relevant to us here have remained the same.[2]

Table 1. Ross's (2002) reconstruction of the PAn verbal paradigm

	Actor	**Patient**	**Location**	**Circumstantial**
INDICATIVE (N>V)				
Neutral	**<um>√**	**√-ən**	**√-an**	**Si-√**
	**k<um>áRaw*	**kaRáw-ən*	**kaRáw-an*	**Si-káRaw*
Perfective	**<umin>√**	**<in>√**	**<in>√-an**	**Si-<in>√**
	**k<um><in>áRaw*	**k<in>aRáw*	**k<in>aRáw-an*	*Si-*k<in>áRaw*
Durative	**<um>REDUP-√**	**REDUP-√-ən**	**REDUP-√-an**	**Si-REDUP√**
	**k<um>a-káRaw*	**ka-kaRáw-ən*	**ka-kaRáw-an*	**Si-ka-káRaw*
NON-INDICATIVE (V)				
Atemporal	**√**	**√-u, √-a**	**√-i**	**án-i+√, √-áni**
	**káRaw*	**kaRáw-u, -a*	**kaRáw-i*	**án-i káRaw, *kaRaw-áni*
Projective	**<um>√-a**	**√-aw**	**√-ay**	**án-ay+√, √-ánay**
	**k<um>aRáw-a*	**kaRáw-aw*	**kaRáw-ay*	**án-ay káRaw, *kaRaw-ánay*

The paradigm termed 'independent' by Wolff (1973) and 'indicative' by Ross (2002) was argued by Starosta et al. (1982/2009) to have come from the reanalysis of participant nominalizations, which were originally used primarily to form rel-

1. Although we demonstrate this change with English, the *-er* suffix has clearly not developed a special role in English relative clauses. On the other hand, the English 'present participle' *-ing* and 'past participle' *-en* do play a role which can be usefully compared to the phenomena under discussion here although space precludes a discussion of this parallel here.

2. I have only added to Table 1 that the indicative forms are hypothesized to have undergone a N > V reanalysis but that the non-indicative forms have remained verbal throughout their history.

ative clauses as well as what could be considered lexical nominalizations. This paradigm still plays a central role in participant nominalization in many modern languages, as exemplified by Paiwan in Table 2, where we see how a single morphological formation plays a dual role as a 'voice marker' as well as a nominalizer.

Table 2. Paiwan voice as thematic nominalization (Ross 2002; Ferrell 1982: 17, 106)

Paiwan	Verb form	Nominalization
k<əm>an	actor voice neutral	'eater', 'someone who eats'
kan-ən	patient voice neutral	'food', 'something to be eaten'
k<in>an	patient voice perfective	'consumed food', 'something eaten'
kan-an	locative voice neutral	'place where one eats'
si-kan	circumstantial voice neutral	'eating utensil', 'something to eat with'

Ross (2002) illustrates the reanalysis of nominalizations as main clause predicates with the Paiwan example in (1), where a possessor in the first stage is reinterpreted as a bona fide agent in the second stage.

(1) Paiwan (Ferrell 1982; Ross 2002)
   a. təkəl-ən a vaua
      drink-PAT.NMLZ SPEC wine
      'the wine is something to be drunk'
   b. təkəl-ən nua qała a vaua
      drink-PAT.NMLZ GEN stranger SPEC wine
      'the wine is the stranger's drinking thing'
      →'the stranger will drink the wine'

We can compare the distribution of the independent paradigm in Puyuma and Tagalog. Both languages require the independent paradigm in relative clauses but in matrix clauses, they diverge. While Puyuma requires a variety of the historical dependent form, Tagalog applies the same independent paradigm used in relative clauses. Most Austronesian languages that preserve these voice paradigms pattern with Tagalog in this respect rather than Puyuma.

(2) Katripul Puyuma (adapted from Teng 2009: 825)
   a. nantu k<in>erutr-Ø
      NOM:3.GEN <PERF.UV>dig-PAT.NMLZ
      'the (thing) dug out'
   b. nantu k<in>erutr-an
      NOM:3.GEN <PERF.UV>dig-LOC.NMLZ
      'the dug out area'
   c. nantu i-kerutr
      NOM:3.GEN CONV.NMLZ-dig
      'the instrument dug with'

(3) Tagalog
   a. ang=h<in>ukay-Ø
      NOM=<BEG>dig-PV
      'the (thing) dug out'
   b. ang=h<in>ukay-an
      NOM=<BEG>dig-LV
      'the dug out area'
   c. ang=i-ni-hukay
      NOM=CV-BEG-dig
      'the instrument dug with'

(4) Nanwang Puyuma (Teng 2008: 47–48)
   a. trakaw dra paisu i isaw
      <AV>steal INDF.OBL money PM.SG.NOM Isaw
      'Isaw stole money.'
   b. tu=trakaw-aw na paisu kan isaw
      3.GEN=steal-PV DEF.NOM money PM.SG.OBL Isaw
      'Isaw stole the money.'
   c. tu=trakaw-ay=ku dra paisu kan isaw
      3.GEN=steal-LV=1SG.NOM INDF.OBL money PM.SG.OBL Isaw
      'Isaw stole money from me.'
   d. tu=trakaw-anay i tinataw dra paisu
      3.GEN=steal-CV PM.SG.NOM his.mother INDF.OBL money
      'He stole money for his mother.'

(5) Tagalog
   a. nag-nakaw si totoy nang pera
      <AV.BEG>steal PM.NOM Totoy GEN money
      'Totoy stole money.'
   b. n<in>akaw-Ø ni totoy ang pera
      <BEG>steal-PV PM.GEN Totoy NOM money
      'Totoy stole the money'
   c. n<in>akaw-an=ako ni totoy nang pera
      <BEG>steal-LV=1SG.NOM PM.GEN Totoy GEN money
      'Totoy stole money from me.'
   d. i-n<in>akaw=niya nang pera ang nanay=niya
      CV-<BEG>steal=3SG.GEN GEN money NOM mother=3SG.GEN
      'He stole money for his mother.'

Noting that Puyuma, Tsou and Rukai do not make use of this paradigm in main clauses, Ross (2009) proposes that the reanalysis of nominalizations as matrix clause predicates defines a phylogenetic subgroup which he terms Nuclear Austronesian. A counter-proposal by Chen (2017) treats the Puyuma split as innova-

tive rather than conservative, specifically, as the result of pattern reduction from a system in which the independent paradigm served to mark a non-indicative mood while the dependent paradigm served as the indicative counterpart. Differential loss resulted in the indicative mood being restricted to main clauses in certain languages and the subjunctive mood being restricted to relative clauses. On this view, the difference between the two paradigms is merely inflectional and does not entail any difference in lexical category; nominalization is merely an illusion. This view has been argued for by Chen (2015); Blust and Chen (2017); Chen and Fukuda (2017) and Jiang (2016) from different perspectives.

An attractive aspect of Starosta et al.'s (1982/2009) hypothesis is that it explains the oddly similar grammatical behavior of the independent paradigm and nouns (Himmelmann 1991; Capell 1964; Scheerer 1924). For instance, the following generalizations appear to hold robustly across the many Philippine languages that have been described to date:

- All independent forms may appear as 'headless' relatives and when apparently 'headed', show no strict pre-head vs. post-head ordering;
- There is no dedicated relativizer that differentiates independent forms in a DP from simple nouns;
- Non-actor voice agents are always expressed identically with possessors
- There are no dedicated copulas that could distinguish between independent forms and simple nouns;
- Independent forms yield lexicalized idiomatic meanings characteristic of derivational morphology.

Kaufman and Chen (2017); Kaufman (2016, 2017, 2009b), building on Starosta et al. (1982/2009), takes other types of evidence to be key, including word order, unusual constituency facts and the well known extraction restrictions that characterize "Philippine-type" languages (Himmelmann 2005). I will not present these arguments here but instead would like to argue from a cross- linguistic perspective that nominalization is so common as to be expected when relative clauses require a special verb form, especially when this form appears to agree with the thematic role or grammatical relation of the relativized argument.[3]

---

3. The theoretical literature abounds with putative parallels between Philippine-type voice and linguistic phenomena of unrelated languages. A partial listing includes Icelandic (Richards 2000), Mayan (Aldridge 2004), Inuit (Kaufman 2009a), Nuer and Kabardian (Chen 2017; Erlewine et al. 2017), Tibetan (Erlewine 2019), Salish (Foley 1998, 2008), Movima (Haude 2025), among others. Here, I focus specifically on the distribution we find in Puyuma, namely, the use of apparent thematic nominalizations to relativize different types of syntactic arguments, which invites comparison to the Cahita languages (Uto-Aztecan) and Galo (Sino-Tibetan).

3. Nominalized "relative clauses": Cross-linguistic parallels

3.1 Undifferentiated nominalization

A large number of languages possess a single morphological marker indicating that a verb serves as part of a relative clause modifier. We review here examples from the Sino-Tibetan, Uto-Aztecan and Turkic families.

Mongsen Ao, despite a wealth of participant nominalizing morphology and a ban on plain finite verbs in relative clauses (Coupe 2007: 80), employs a general nominalizer *-pàʔ* for the lion's share of relativization (Coupe 2007: 219), as shown in (6).

(6) Mongsen Ao (Coupe 2007: 219–222)

a. AGENT RELATIVE
[[a-ji tʃəm-ja li-pàʔ] a-miʔ tʃu]$_{NP}$
NRL-rice.beer drink-CONT be-NMLZ NZP-person DIST
ke-nu-ùʔ
1S.POSS-yngr.sibl-DECL
'The guy drinking the rice beer is my younger brother.'

b. PATIENT RELATIVE
[[mətʃatshəŋ nə tsəŋ-pàʔ] a-úk sə]$_{NP}$
Mechatseng AGT spear-NMLZ NZP-pig ANAPH
'The pig that Mechatseng speared…'

c. TEMPORAL/OBLIQUE RELATIVE
[[sàti tʃhà-pàʔ] hmapaŋ]$_{NP}$ ku
marriage do-NMLZ time LOC
'At the time of marrying, …'

In Pima Bajo, an Uto-Aztecan language, a single nominalizer with an origin in a participial and demonstrative functions to relativize both subjects and non-subjects (Estrada-Fernández 2008, 2012), as seen in (7). Note that agents of non-subject relatives are expressed as genitive possessors in Pima Bajo (glossed below as NSBJ).

(7) Pima Bajo (Estrada-Fernández 2008: 73)

a. gogos kɨia kil [hink–im-k=ɨg]
dog bite.PERF man shout-IMPF-REL
'The dog bit the man that was shouting.'

b. okosi [in=nɨid-kɨg] nɨʔi-im
woman 1SG.NSBJ=see-REL sing-IMPF
'The woman I saw.'

Turkic languages are well known for using a wide range of participles for both relativization and tense-aspect functions. Some of these languages, like Karachay-Balkar have a general participle form that is used for both subject and non-subject relatives, as seen with *-ɣan* in (8).

(8) Karachay-Balkar (Comrie 1998: 81)
a. [kitab-i al-ɣan] oquwču
book-ACC buy-PRT student
'the student who bought the book'
b. [oquwču al-ɣan] kitap
student buy-PRT book
'the book that the student bought'

3.2 Two-way subject vs. non-subject differentiation

A wide array of languages also impose a two-way subject/non-subject or agent/non-agent distinction in the formation of relativized predicates. Below, we briefly review the relevant facts in languages related to those presented in Section 3.1, specifically: Dolakha Newari (Sino-Tibetan), Turkish (Turkic), and Ute (Uto-Aztecan), as well as Quechua (Quechuan).

Dolakha Newari (Genetti 2007), another Tibeto-Burman language, displays a basic opposition between *-gu* and *-a* in subject and non-subject relatives, as in (9) and (10), respectively.

(9) Dolakha Newari (Genetti 2007: 312)
SUBJECT NOMINALIZATION
[pali depān coŋ-gu]$_{REL}$ kok
roof on stay-SBJ.NMLZ crow
'the crow that is on the roof'

(10) Dolakha Newari (Genetti 2007: 313)
NON-SUBJECT NOMINALIZATION
a. [jin phoŋ-a]$_{REL}$ misā
1S.ERG ask.for-NSBJ.NMLZ woman
'the woman whom I asked for (in marriage)…'
b. [thamun he yeŋ-a]$_{REL}$ ṭhɜ̃ĩ
2HON.ERG work do-NSBJ.NMLZ place
'The place where you work…'

Turkish also shows two-way opposition between subject oriented *-en* and non-subject oriented *-diğ* participles, as seen in (11a) and (11b). This two-way split dis-

tinguishes Turkish from the related Karachay-Balkar, seen earlier in (8), which had but a single all-purpose relativizer, *-ɣan*.[4]

(11) Turkish (Comrie 1998: 82)
a. [kitab-ı al-an] öğrenci
book-ACC buy-SBJ.PRT student
'the student who bought the book'
b. [föğrenci-nin al-dığ-ı] kitap
student-GEN buy-NSBJ.PRT-3.POSS book
'the book which the student bought'

Note that, as with all the other languages examined here, the participle may serve as an adnominal modifiers, as in (12a), or stand alone as the apparent head of the relative clause, as in (12b). The same holds true for subject relatives, as seen in (13).

(12) Turkish (Kornfilt 1997: 63)
a. [adam-ın ye-diğ-i] balık
man-GEN eat-NSBJ.PRT-3.POSS fish
'the fish that the man eats/ate'
b. [adam-ın ye-diğ-i]
man-GEN eat-NSBJ.PRT-3.POSS
'what the man eats/ate'

(13) Turkish (Kornfilt 1997: 63)
[para-yı ver-en]
money-ACC give-SBJ.PRT
'(the person) who gave the money'

Turkish furthermore applies genitive case to the agent of the non-subject participle, as seen in almost all the languages examined here. The parallelism between possessors and object participle agents can be seen in (14a) and (14b), respectively, where the agent/possessor dependent is marked with genitive case and the head displays possessor agreement.

(14) Turkish (Krause 2001: 31)
a. Hasan-ın odas-ı
Hasan-GEN room-3.POSS
'Hasan's room'

4. Hankamer and Knecht (1976) establish that grammatical function is key to the choice of participle in Turkish, as opposed to linear order or other considerations (cf. Underhill 1972), although other factors come into play, as well, for instance, the specificity of the subject (Kornfilt 1997: 59).

b. Meltem-in gör-düg-ü yılan
Meltem-GEN see-NSBJ.PRT-3.POSS snake
'the snake that Meltem saw'

Note also that the participle constructions in Turkish show evidence of clause reduction or at least the absence of higher functional categories associated with main clause predication. In particular, they lack complementizers, relative pronouns and tense, although they often do contain aspectual distinctions.

Despite a good deal of dialectal variation, Quechua also shows a two-way distinction in nominalized relatives, marked by a reflex of Proto-Quechua *-*shqa* for non-subjects and -*q* for subjects, as seen in (15a) and (15b).

(15) Huallaga Quechua (Weber 1989: 281)
a. [Runa maqa-**sha**-yki] sha:-yka-mu-n
man hit-NMZR-2POSS come-IMPF-DIST-3
'The man whom you hit is coming.'
b. [Runa maqa-shu-**q**] sha:-yka-mu-n
man hit-2POSS-SBJ.NMZR come-IMPF-DIST-3
'The man who hit you is coming.'

Similar to Turkish, Quechua shows genitive case on the agents of non-subject nominalization, which also lack a relativizer or relative pronoun. Gijn (2024) notes that it is not only Quechua but in fact all Central Andean languages that employ nominalization in a relativization function. Lehmann (1984) shows that these forms are used both productively to form relative clauses, as in (16a), but are also found in lexicalized nominalizations which may possess idiosyncratic meanings, as in (16b).

(16) Ancash Quechua
a. miku-na-ː-ta muna-nki-ku?
[eat-NMLZ.POT-1SG]-ACC want-2-QM
'Do you want what I will eat?'
b. puklla-na
play-NMLZ.POT
'toy'

If an adpositional argument is relativized, the relevant postposition takes a nominalization formed with the non-subject nominalizer as its complement, seen in (17). We will return to Quechua in more detail below in Section 4.4 to consider the case marking pattern in relatives.

(17) Cuzco Quechua (Lefebvre & Muysken 1988: 198)
a. Riqsi-ni [may uruma-sqa-yki-pi]
know-1 where fall-NMLZ-2-LOC
'I know (the place) where you fell.'
b. [Xwancha-q rima-sqa-n-wan] riqsi-nki-chu?
Juan-GEN speak-NMLZ-3-with know-2-QM
'Do you know (the person) with whom John talked?'

The last language to be considered here is Ute, whose relative clauses have been discussed at length by Givón (2011, 2012). A plain declarative main clause is shown in (18). Ute, like Turkish and Quechua but unlike Newari, shows a nominative-accusative case marking pattern, which is expressed through a variety of morphological mechanisms, including final vowel devoicing (Givón 2011).

(18) Ute (Givón 2011: 97)
mamachi̥ sivaatuchi paqha-vaa-ni
woman/SBJ goat/OBL kill-IRR-FUT
'The woman will butcher the goat.'

Subject relative clauses, as in (19a), are formed with the dedicated nominalizers *-tʉ/-rʉ*, while object relatives are formed with the object oriented nominalizer *-na*, as shown in (19b). As seen in this latter example, the agent must now take the genitive case, just as in the languages examined previously.

(19) Ute (Givón 2011: 97, 278)
a. SUBJECT RELATIVE CLAUSE
'áapachi̥ 'u [sivaatuchi 'uway pa̧qha-pʉga-tʉ]…
boy/SBJ 3S/SBJ goat/OBJ 3S/OBJ kill-REM-SBJ.NMLZ
'the boy who killed the goat…'
b. OBJECT RELATIVE CLAUSE
sivaatuchi̥ [mamchi paqha-vaa-na]…
goat/SBJ woman/GEN kill-IRR-OBJ.NMLZ
'the goat that the woman will butcher'

Oblique and adpositional relatives also employ the object nominalizer *-na*, in conjunction with a relativizer *pʉ*, which hosts whatever additional postpositions are required. This is seen in (20), where the relativized argument is a comitative. The verb, *wʉ́ʉka* 'work', is nominalized by means of *-na* and the relativizer *pʉ* hosts the postposition *-wa* 'with'. This represents an unusual split system with a case marked relativizer that has a structural analogue in many European languages, as well as the nominalization strategy found commonly elsewhere.

(20) Ute (Givón 2011: 282)
mamchị ˈu [pʉ-wa taˈwachi wʉ́ʉka-qha-nạ
woman/SBJ the/SBJ RELT-with man/GEN work-ANT-OBJ.NMLZ
'the woman with whom the man worked…'

As Givón exemplifies in (21), apparent headless relative clauses are common in Ute. Their independence as stand alone arguments is most likely linked to their basis in nominalizations.

(21) Ute (Givón 2011: 288)
HEADLESS RELATIVE CLAUSE
[ʼuru ʼáapachi ʼuway tʉka-pʉga-nạ]…
that/OBL boy/GEN the/GEN eat-REM-OBJ.NMLZ
'what the boy ate…'

3.3 Three and four-way differentiated systems

While the languages surveyed in the preceding subsection display a basic two-way opposition between subject and non-subject nominalization for the purpose of relative clause formation, other languages have further elaborated this pattern so that non-subject relativization may come in two or three different varieties itself. A typical example is described by Marlett (2012: 218) for Seri, a language of northwestern Mexico which possesses three participant oriented deverbal forms used in relative clauses (RCs):

> RCs in Seri all utilize deverbal nouns, which come in three types: deverbal noun oriented toward the subject, regardless of its semantic role (DN.S), deverbal noun oriented towards the direct object (DN.O), and deverbal noun oriented toward other nominals, including indirect objects and various kinds of obliques (DN.I). The choice between different deverbal nouns depends on the grammatical relation that the head noun has in the RC. Thus DN.S is used when the head noun corresponds to the subject of the RC, DN.O when it corresponds to the direct object (whether primary object or secondary object, to use the terminology of Dryer 1986), and DN.I when it is none of the above.

Marlett (2012: 216–217) further notes that the agents of non-subject relatives are expressed as possessors and that the language lacks relative pronouns and complementizers, as is the case for most of the languages reviewed here.

In (22), we see a typical two-way distinction in Tibetan between a subject and object relative clause, the former whose verb is nominalized with *=mkhan* and the latter, which is nominalized with *-pa*.

(22) Central Tibetan (DeLancey 1999)
   a. mi sha gtub=mkhan de
      person meat chop=NMLZ1 DEM
      'the person who chopped the meat'
   b. nga-s sha btsos-pa de
      1SG-ERG meat cook-NMLZ2 DEM
      'the meat which I cooked'

In fact, Tibetan shows a richer inventory of nominalizers used to form relative clauses and, most significantly, we can trace the etymologies of many of these morphemes to their lexical origins, as seen in (23). In most cases, the forms even continue to lead a dual life, both as bound nominalizers as well as free lexemes, continuing their pre-grammaticalized function.

(23) TIBETAN NOMINALIZERS/RELATIVZIERS
   a. AGENT.NMLZ/RELT *-mkhan* < 'expert'
   b. AGENT.NMLZ/RELT *-mi* < *mi 'person, man'
   c. NMLZ/RELT *-pa* < *pa 'father'
   d. LOC.NMLZ/RELT *-sa* < *sa 'earth, ground, place'

With regard to how these lexemes became grammaticalized as participant nominalizers, Cristofaro (2019) notes:

> As repeatedly pointed out in the literature (DeLancey 1986, 1999, LaPolla & Huang 1996, Noonan 1997, Gildea 1998, Givón 2012, among others), a possible source for nominalized relative clauses are constructions where the lexical items that ultimately give rise to the nominalizer occur themselves as heads of a relative clause, e.g. 'the person who likes drinking liquor', 'the place that I came from'.

Highly animate nouns (e.g. 'person'), as canonical doers, may take on an agentive function, while inanimate nouns may take on patient-like or objective functions. More transparently, nouns like 'place' come to function as relative markers for locations and perhaps oblique arguments more generally. Tibetan plays an especially important role in helping us understand the diachronic origins of highly differentiated relativization marking. In related languages, the synchronic link between nominalization and relativization is clear but not all of the morphemes involved can be etymologized as clearly as in Tibetan.

In Japhug, another Tibeto-Burman language, relativization requires nominalized participles, which Jacques (2021: 713) describes as follows:

> Participles are nominalized verb forms that keep some verbal characteristics: they can serve as predicates of subordinate clauses (in particular relative and complement clauses), take TAM, polarity and associated motion marking, and preserve the verb's argument structure.

On the other hand, as Jacques (2021:713) notes, nominalized participles cannot serve as the predicate of a main clause, nor do they take the complex person marking morphology associated with verbs, perfectly parallel to Puyuma, as reviewed earlier in (2)–(4). Instead, person is marked by possessive pronouns, which index a single argument in the clause.[5] Examples of Japhug nominalization/relativization are shown for subjects in (24), for objects in (25), and for locatives in (26). Note that locative relativization can also apply to dative arguments, as shown in the second example in (26).[6]

(24) Japhug (Jacques 2016)
SUBJECT RELATIVIZATION
a. kɯ-si
NMLZ:S/A-die
'The dead one.'
b. ɯ-kɯ-sat
3S.POSS-NMLZ:S/A-kill
'The one who kills him.'

(25) Japhug (Jacques 2016)
OBJECT RELATIVIZATION
a. kɤ-sat
NMLZ:P-kill
'The one that is killed.'
b. a-kɤ-sat
1SG.POSS-NMLZ:P-kill
'The one that I kill.'

(26) Japhug (Jacques 2016, 2021)
LOCATIVE RELATIVIZATION
a. ɯ-sɤ-ɣi
3SG-NMLZ:OBL-come
'The place/moment where/when it comes.'

5. Incidentally, Japhug, which is an ergative language, also allows of relativization of absolutive undergoers using a finite clause, in addition to the nominalization strategy required for other arguments. It shares this property with Chamorro, an Austronesian language which employs a participle strategy for most arguments but also allows relativization of absolutive undergoers using a plain finite clause (Chung 1998).

6. The nominalizing morphemes in question are glossed as NMLZ in Jacques (2016) but as PCP (participle) in Jacques (2021). S/A refers to the identification of the nominal with the sole argument of an intransitive verb or the agentive argument of a transitive verb while P refers to identification with the undergoer of a transitive verb.

b. a-sɤ-tʰu
1SG.POSS-OBL:PRT-ask
'The one whom I ask'

The Tani languages Apatani and Galo come the closest to conservative Austronesian languages in showing a four way distinction in nominalized relatives; regarding this subgroup, Post and Modi (2011) note, "In all Tani languages for which adequate data exist (basically, Galo, Minyong, Mising and to an extent Apatani and Tagin), relative clauses are based on predicate nominalizations which reference grammatical relations." Apatani's inventory of nominalizers, all of which are obligatorily used in relativization, can be seen in (27) (Abraham 1985; Krause 2001).

(27) APATANI NOMINALIZERS/RELATIVIZERS
a. AGENT RELATIVIZATION *-bo*
b. PATIENT RELATIVIZATION *-ni*
c. LOCATIVE RELATIVIZATION *-ko*
d. INSTRUMENTAL RELATIVIZATION *-nanɨ*

Not only are the semantics of the thematic nominalizers in (27) apparently identical to those of Austronesian, we also find a cluster of shared grammatical properties, as well. In particular, Krause (2001) points out that Apatani lacks complementizers, relative pronouns and employs these predicates as 'verbs with nominal properties', including genitive case subjects.[7] Furthermore, the relativized verb can serve as the head itself, all features which are very familiar from "Philippine-type" languages of the Philippines and surrounding areas. The examples in (28) show how the forms in (27) are deployed in relative clauses.

(28) Apatani relative clauses (Abraham 1985)
a. [Alyi-mi ka lanɨ-bo nyɨmɨ] hɨ aya do
pig-ACC GEN catch-AGT.NMLZ woman DET good EXT
'The woman who caught the pig is good.'
b. Mólu [Kago-ka ta-nɨ my]-mi kapabɨne
they Kago-GEN kick-PAT.NMLZ man-ACC see:PST
'They saw the man whom Kago kicked.'

7. The genitive case on relative clause agents is common to the other languages discussed here but the genitive case stacked upon the accusative marked object in the first example of (28) is not. In regard to this, Abraham (1985:131) states, "When the subject noun phrase is relativized, 'ka' — the genitive marker — occurs with the object noun phrase in addition to the accusative marker already present with it."

c. [Mó-ka iɲ-ko lemba]-mi ɲo kapato
3SG-GEN go-LOC.NMLZ village-ACC 1SG see:PST
'I saw the village where he went.'
d. [ɲi-ka pani-nani ilyo]-mi mó biti
1SG-GEN cut-INSTR.NMLZ sword-ACC 3SG bring-PST
'He brought the sword with which I cut.'

Each relativizer above is described by Abraham (1985: 119) as forming lexical nominalizations, as well, e.g. *ini-nani* (walk-INSTR.NMLZ) 'walking stick'; *ipi-nani* (sweep-INSTR.NMLZ) 'broom'.[8] Galo, another Tani language, employs the same strategy with a set of partially related morphemes, as discussed by Post (2011) and shown in (29).

(29) Galo relative clauses (Post 2011)
a. ŋó [dór-rɪ́ pá-tà-nà]$_{RELC}$=əə
1SG CLF:high.animal-ten chop-INCP-NMLZ:SBJ=COP.IMPF
'I am (he) who is to sacrifice ten (mithuns).'
b. higì [[ŋó-kà dó-há]$_{RELC}$ (jaràa)]$_{NP}$=əə
this 1SG-GEN eat-NMLZ:NSBJ.IRR (goods)=COP.IMPF
'This is what I'll eat.'
c. [[nó-kə agóm takàa-nam]$_{RELC}$ gadə]$_{NP}$
2SG-GEN speech ask-NMLZ:NSBJ.RLS group
'the bunch of questions that you asked'

Finally, returning to the Uto-Aztecan languages, we have already seen in (7) above that some members of this family, such as Pima, have a single nominalization used in relative clauses of all types. We have also seen that other members of the family, like Ute, have a two-way distinction between subject and non-subject nominalization in relative clauses. To complete the picture, we observe that the Cahita subgroup of this family uses distinct nominalizers for agents (or proto-agents), patients (or proto-patients) and locatives, as seen in (30), from Yaqui.

(30) Yaqui (Álvarez González 2012: 72,73; Álvarez González 2016: 130)
a. AGENT RELATIVIZATION
Joan uka chu'u-ta [Maria-ta ke'e-ka-m]-ta me'a-k
John DET.ACC dog-ACC Mary-ACC bite-PERF-AGT.NMLZ-ACC kill-PERF
'John killed the dog that bit Mary.'

8. Note that some nominalizations, like the instrumental, are complex and contain a more general nominalizer *-ni* in addition to a unique marker (e.g. *-na* for the instrumental, *-bo* for the agent).

b. PATIENT RELATIVIZATION
U bisikleeta [in jinu-ka-'u] sikili
DET bicycle 1SG.GEN buy-PERF-NSBJ.NMLZ red
'The bicycle that I bought is red.'

c. LOCATIVE RELATIVIZATION
Káha-ta-e'e nee-míika [póhporo-m em é'ekria-'apo]
box-ACC-IMP 1SG.ACC-give match-PL 2SG.GEN keep-LOC.NMLZ
'Give me the box where you keep the matches.'

Álvarez González (2012, 2016) shows that the markers employed in relative clauses also function as typical nominalizers, forming noun phrases of the type shown in (31).

(31) Yaqui (Álvarez González 2012: 82)
a. ye-ye'e-me
RED-dance-AGT.NMLZ
'dancer'
b. tekipanoa-me
work-AGT.NMLZ
'worker'

Note also that, in Yaqui, just as in nearly all the languages reviewed here, the agents of non-subject nominalizations are expressed identically to possessors, as was true for Old Cahita, the parent language of Yaqui, Mayo and Tehueco (Álvarez González 2016: 108), as shown in (32).

(32) Old Cahita (Álvarez González 2016: 123)
ioreme [em-veb-tevo-ye]…
person 2SG.GEN-beat-CAUS.INDIR-NMLZ
'The person that you order to beat…'

Finally, note that like Austronesian and many of the languages reviewed earlier, Old Cahita and its descendants employ participant nominalizers in conjunction with aspect morphology. The basic elements of actor and patient nominalizations and their layering can be compared between Old Cahita and Proto-Malayo-Polynesian (following the widely accepted reconstruction of Ross 2002). Comparing first the agent nominalization of Old Cahita with the actor voice of PMP, we find that while the linear ordering differs the layering of elements is the same; aspect is internal to the nominalization, as seen in (33) and (34).

(33) Old Cahita
a. √eria-me
love-AGT.NMLZ
'the one who loves'

b. √eria-ca-me
love-PERF-AGT.NMLZ
'the one who loved'
c. √eria-naque-me
love-FUT-AGT.NMLZ
'the one who will love'

(34) PMP (approximation)
a. <um>√ibeR
<AGT>love
'the one who loves'
b. <um><in>√ibeR
<AGT><PERF>love
'the one who loved'
c. <um>i~√ibeR
<AGT><IMPF>love
'the one who will love'

When we compare the patient nominalization of Old Cahita to the Austronesian patient voice, we now see the positioning of the genitive agent outside the nominalization, as in (35) and (36).

(35) Old Cahita
a. in-√eria-ye
1SG.GEN-love-PAT.NMLZ
'The one that I love'
b. in-√eria-ca-ye
1SG.GEN-love-PERF-PAT.NMLZ
'the one that I loved'
c. in-√eria-naque-ye
1SG.GEN-LOVE-FUT-PAT.NMLZ
'the one that I will love'

(36) PMP (approximation)
a. √ibeR-en=ku
love-PAT.NMLZ=1SG.GEN
'The one that I love'
b. in-√ibeR-Ø=ku
PERF-love-PAT.NMLZ=1SG.GEN
'the one that I loved'
c. i~√ibeR-en=ku
IMPF-love-PAT.NMLZ=1SG.GEN
'the one that I will love'

Putting aside linear order, which may be determined independently, all the orders above are consistent with the hierarchical layering schematized in (37).

(37) POSS[NMLZ[ASPECT[[√]]]]

Therefore, beyond the mere appearance of differentiated participant nominalizers used to relativize various arguments (and adjuncts), we find a similarly layered structure and a group of shared typological features, which are explicated further in Sections 3.4 and 4 below.

3.4 The typology of nominalized relatives

We can define the nominalized relative prototype as displaying the characteristics in (38), some of which have already been discussed above (a)–(d), and others (e), (f), which we will turn to below (cf. Krause 2001; Ackerman & Nikolaeva 2013).

(38) THE NOMINALIZED RELATIVE PROTOTYPE
- a. multiple thematically oriented nominalization/relativization markers
- b. possessor/genitive case for agents and possibly other arguments
- c. no dedicated complementizer or relative pronoun (the "gap strategy")
- d. no fixed order between the 'relative clause' and the 'head noun'
- e. rampant 'headless relatives'
- f. multiple determiner phrases within an argument phrase

An analysis rooted in (derivational) participant nominalization has the power to explain the full constellation of facts above. First, we have seen that participant nominalizations come in many varieties, targeting agents, patients, locations and instruments, as well as grammatical relations, such as subject and object. On the other hand, there are no clear examples of equally multifarious voice (and the applicative analysis of Philippine-type voice faces its own problems, see Kaufman 2017). Second, with regard to (38b), it is well known that arguments, and especially agents, are typically expressed as possessors in nominalization (Koptjevskaja-Tamm 1993; Malchukov 2004), thus accounting for the pervasive ergative-genitive homophony in clause types that have developed from nominalizations (Manning 1996; Kaufman 2009a).

The subsequent four features, (38c)–(38f), all follow from the grammaticalization of two noun phrases in apposition to a subordinate relative clause, as described by Comrie and Thompson (1985); Givón (2001, 2011, 2012); Shibatani (2009, 2019).[9] Specifically, if the source construction is based on simple apposi-

9. On the path from apposition to relative clause, Comrie and Thompson (1985: 394) note:"[I]t is not difficult to understand how a nominalization can function as a relative clause: the nom-

tion, we do not expect any markers of subordination (38c), nor do we expect fixed order between the head and the co-referring nominalization, as they are in a symmetrical relation (38d), nor do we expect the head to be obligatorily present (38e), as the nominalization can occupy all the positions and fulfill all the functions that a simple noun may fulfill.

Finally, if the apposition structure contains a full DP, we would expect doubling of determiners, case markers, or any other elements on the functional periphery of argument phrases.

We can again turn to the Cahita languages (Álvarez González 2012, 2016, 2019) to illustrate this diachronic pathway. Álvarez González (2012, 2016, 2019) illustrates the path from apposition to subordination with the synchronic possibilities in (39).

(39) Yaqui (Álvarez González 2019: 306)

a. APPOSITIVE
U yoeme, (u) tekipanoa-ka-me, naamukia
DET man DET work-PERF-AGT.NMLZ drunk
'The man, the one who worked, is drunk.'

b. RELATIVE CLAUSE
U yoeme tekipanoa-ka-me naamukia
DET man work-PERF-AGT.NMLZ drunk
'The man who worked is drunk.'

Whereas in (39a), interpreted as a non-restrictive relative, there is a prosodic break between *u yoeme* and *u tekipanoakame*, as well as the possibility of a determiner for both coreferent phrases, in (39b), there is no prosodic break, no possibility of a determiner before the apparent relative, and the structure is interpreted as a restrictive relative. Returning now to Puyuma, where we began, we find that modification is expressed by juxtaposition of a case marked phrase regardless of whether the modifier is a simple adjective, as in (40a), or an apparent relative clause, as in (40b) and (40c). In other words, one of the crucial ingredients in the source structure posited by Álvarez González (2012 et seq) for Yaqui, is not only possible but obligatory in Puyuma and, as we will see, other Formosan languages, as well.

---

inalization and the noun with which it is in construction can be thought of as two juxtaposed nominal elements [NOM] [NOM], the modifying relationship between them being inferred by language-users (rather than being specified by the grammar, as it is in languages with specific relative clause morphology), just as the modifying relationship is inferred in a noun-noun compound such as tree-house, in which the two nominal elements simply happen to be single nouns."

(40) Puyuma (Teng 2008: 54)

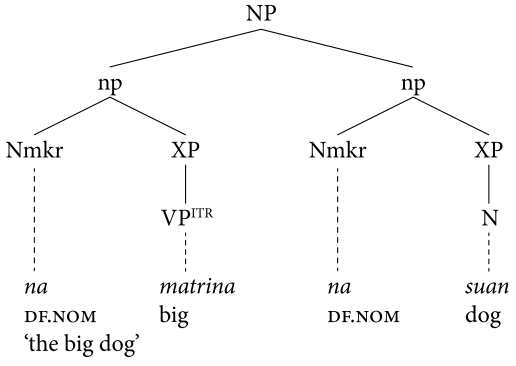
a. tu=alrak-aw na kiping na bulray
3.GEN=take-TR1 DEF.NOM clothes DEF.NOM beautiful
'They took the beautiful clothes.'

b. na ma-ra-rengay na trau
DEF.NOM ITR-RED-tell DEF.NOM person
'the person who is talking'

c. nanu i-veray na valray mu, sahar=ku
DEF.NOM.2S.GEN NMLZ:CV-give DEF.NOM book TOP like=1SG.NOM
'The books you gave, I like them.'

As Teng (2008: 103) notes, "In English, attributive constructions and relative clauses are structurally different, but this distinction does not occur in Puyuma." Teng analyzes the relation between NP-internal phrases in a modification relation as a symmetrical one, with each sub-phrase constituting an NP, subsumed under the NP argument phrase, as in (41).

(41)



As expected from a symmetrical relationship, the ordering of np constituents within the NP (with the exception of demonstratives) is free, as shown by (42). Note that multiple determination is required regardless of whether the modifiers are relative-like, as seen earlier, or bare modifiers, as seen here.

(42) Puyuma (Teng 2008: 105)

a. [na suan] [na matrina] [na utreutrem]
DEF.NOM dog DEF.NOM big DEF.NOM black

b. [na matrina] [na utreutrem] [na suan]
DEF.NOM big DEF.NOM black DEF.NOM dog

c. [na matrina] [na suan] [na utreutrem]
DEF.NOM big DEF.NOM dog DEF.NOM black
'the big black dog(s)'

Case matching is also required between np constituents within an NP, as shown in (43).

(43) Puyuma (Teng 2008: 82)
a. [na suan] [na matrina] [na utreutrem]
DEF.NOM dog DEF.NOM big DEF.NOM black
'the big black dog(s)'
b. me-na'u=ku [dra matrina] [dra utreutrem] [dra suan]
ITR-see=1SG.NOM INDF.OBL big INDF.OBL black INDF.OBL dog
'I saw a big black dog.'

The use of multiple determiners in modification structures is found in several Formosan languages, including Saaroa, shown in (44), and Budai Rukai, shown in (45). As these examples show, multiple determiners are required not only in structures that translate to relative clauses but also those that might translate to compounds or simple modification, as in (45b).

(44) Saaroa (Pan 2012: 289)
uka'a=cu ka ma-m-a-ini [ka m-ari-vakese
NEG=COS.ASP CORE RED-AV-STAT-small LNK AV-hand.motion-BOUND.ROOT
tasau]
dog
'The child who beat a dog is gone.'

(45) Budai Rukai (Sung 2011: 532; Jiang 2016)
a. ngu-a-bere [ku wa-lrumay ki lrailrai] [ku lasu]
go-RLS-escape NOM RLS-beat OBL L. NOM man
'The man that beat Lrailrai ran away.'
b. wa-kane=aku [ku lrubu] [ku lapanay]
RLS-eat=1SG.NOM OBL porridge OBL corn
'I ate corn porridge.'

The distribution of this pattern across Rukai, Puyuma and Saaroa argues strongly for its reconstructability to Proto-Austronesian, as all higher-level subgrouping hypotheses with the exception of Sagart (2004) take these to represent three independent branches of Proto-Austronesian. The import of this fact has never been explored but I would argue that it represents the original *motivation* for nominalization in Austronesian relatives. If subordination, or at least subordination within NP, was highly constrained so that all modification was expressed by apposition and if bona fide verbs could not appear embedded within an NP/DP, as is often the case, then we can predict the use of participant nominalization for the purpose of relativization.[10] The Austronesian evidence thus suggests that it is the strong con-

10. Here and throughout I employ the term relativization loosely despite the apposition structure I refer to not being a relative clause in any strict sense. See Shibatani (2019) and Álvarez González (2019) for further discussion.

straints on subordination which lead to nominalized relatives rather than the mere existence of participant nominalizations that give rise to apposition.[11]

In the following section, we show how the forms of the participant nominalizers themselves in Austronesian can be explained by seemingly parallel developments in the Cahita and Tani languages.

4. Case marking and participant nominalization

4.1 The Cahita languages

As discussed by Álvarez González (2012 et seq), Old Cahita, the Uto-Aztecan antecedent of Yaqui, Mayo and the no longer spoken Tehueco language, displayed the nominalizers shown in (46).

(46) OLD CAHITA NOMINALIZERS (Álvarez González 2012)

a.	SUBJECT NOMINALIZER	*-me, -u*
b.	NON-SUBJECT NOMINALIZER	*-ye*
c.	LOCATIVE NOMINALIZER	*-ye-po* (NON-SUBJECT+LOCATIVE)

Álvarez González (2012: 131) notes that the diachronic source for two of the Cahita participant nominalizers appears to be found in *case marking postpositions*,[12] stating:

> We can observe that the sources of the patient and locative nominalization suffixes in Cahita seem to be postpositional. The patient nominalization suffix *-ye* from Old Cahita was indeed related to the instrumental postposition, whereas the patient nominalization suffixes *-'u* and *-'wi* come from the directional postposition in Yaqui and Mayo…

11. This rationale stands as an alternative to Aldridge (2023), who proposes that nominalization was a strategy for extraction of an argument across a nominative phrase. Aldridge's view explains why agent relativization, as opposed to non-agent relativization, does not appear to have required a special nominalization in PAn and most of its descendants. However, it does not explain why many languages that pattern ergatively (e.g., Puyuma, Tibetan) still require nominalization for patient extraction (but see Footnote 7 for a caveat to these exceptions) nor does it explain the use of nominalization for subject relativization in a number of a non-Austronesian nominative-accusative languages (e.g., Cahita, Tani, Quechua). It is not clear at this point if the nominalization strategy for relativization skews towards nominative-accusative languages or whether the attested distribution is expected on statistical grounds, but this is an important area for further exploration.

12. In subsequent work, Álvarez González (2019, 2023) traces the origin of both the instrumental and the nominalizer functions of *-ye* to a marker of inalienable inanimate possession in Proto-Uto-Aztecan. For present purposes, we can put this longer view aside to focus on the connection between the nominalizer and the postpositional case marker.

An example of the instrumental use of *-ye* in Old Cahita compared with a non-subject nominalization derived by the same morpheme is shown in (47). The same is shown for the directional in the modern Cahita language, Yaqui, in (48).

(47) Old Cahita (Álvarez González 2016: 123)

a. POSTPOSITIONAL FUNCTION
cuta-ye nee mea-c
wood-INST 1SG.ACC hit-PERF
'He hit me with a stick.'

b. NON-SUBJECT RELATIVIZATION
ioreme [em-veb-tevo-ye]…
person 2SG.GEN-beat-CAUS.INDIR-NSBJ.REL
'The person that you order to beat…'

(48) Yaqui (Álvarez González 2012: 72,73)

a. POSTPOSITIONAL FUNCTION
Navojoa-u ne-siika-k
Navojoa-DIR 1SG.NOM-go-PERF
'I went to Navojoa.'

b. NON-SUBJECT RELATIVIZATION
U bisikleeta [in jinu-ka-'u] sikili
DET bicycle 1SG.GEN buy-PERF-NSBJ.REL red
'The bicycle that I bought is red.'

The adpositional use of the Old Cahita locative is shown in (49) and the dual use of the cognate suffix in modern Yaqui is shown in (50).

(49) Old Cahita
baa-po
water-LOC
'into the water'

(50) Yaqui (Lindenfeld 1973: 17)

a. POSTPOSITIONAL FUNCTION
gwaimam-po hente yoi-nooka
Guaymas-LOC people Spanish-speak
'In Guaymas people speak Spanish.'

b. LOCATIVE RELATIVIZATION
Wa kari [nim bo'e-pea-'apo] ujyooli
DEM house 1SG.GEN sleep-DES-LOC.REL pretty
'That house that I want to sleep in is pretty.'

Although questions remain, the link in the Cahita languages between postpositions/case markers on the one hand and the corresponding participant nominal-

izers on the other is clear. We now turn briefly to similar evidence from the Tani languages of Northeast India, also introduced earlier in Section 3.3.

4.2 The Tani languages

The full inventory of primary participant nominalizers in Galo is given in (51).

(51) GALO PRIMARY NOMINALIZERS (Post 2011)

a.	SUBJECT NOMINALIZER	*-nà*
b.	REALIS EVENT AND NON-SUBJECT CORE NOMINALIZER	*-nam*
c.	IRREALIS/OBLIGATIVE EVENT AND NONSUBJECT CORE NOMINALIZER	*-há*
d.	LOCATIVE/OBLIQUE NOMINALIZER	*-kò*

With regard to the diachronic sources for these nominalizers, Post (2011: 280) states:

> Although it is not currently possible to reconstruct Proto-Tani etyma to pre-Proto-Tani levels, it may not be unreasonable to speculate that a pre-PT subject or agentive nominalizer *na could relate to agent-oriented articles/nominalizers or relational markers such as Athpare *-na* 'Article/singular nominalizer' (Ebert 1997; Bickel 1999) or Mongsen Ao *nə* 'Agentive case marker' (Coupe 2007); event/non-subject nominalizer *nam might then reflect a subsequent fusion with PT object marker *m, as *na=m. Locative/oblique nominalizer *-kò* has a plausible though uncertain partial cognate in a nominal root *kó-* 'place; ground'…

The object marking function Post posits here for the *-m* portion of (51b) is exemplified in (52), where *isì* 'water' takes the accusative case marker assigned by the transitive verb *círdùu* 'boil'.

(52) Galo (Post 2011)
[ŋó]$_A$ [isì=əəm]$_O$ [cır-dùu]$_{PRED.TR}$
1SG water=ACC boil.VT-IMPF
'I'm boiling the water.'

While there are competing possibilities for the etymologies of the nominalizers in (51) (Mark Post, p.c. 5/22/25), it is highly suggestive that the nominalizer used for objects can be analyzed as a nominalizer combined with an object case marker, especially given the Cahita evidence reviewed above.

4.3 Early Austronesian case suffixes

The apparent correspondences between case/adpositional functions and participant nominalizers in the Cahita and Tani languages is especially remarkable in light of an underappreciated fact about the Austronesian voice system, namely that the patient voice and locative voice suffixes appear on pronouns and animate NPs as accusative and oblique case markers, respectively. The *-an suffix seen earlier in Table 1 and best known for its locative voice function, is reconstructed by Ross (2006: 528) as a PAn locative (case) suffix with evidence from Proto-Atayal, Siraya, Proto-Amis, Proto-Rukai and Kavalan.[13]

Similarly, Ross (2006) reconstructs accusative pronouns on the basis of the comparative evidence seen in Table 3. The place of the accusative pronouns in the entire paradigm is shown in Table 4. The perfect phonological correspondence between both the patient nominalizer suffix *-en and the accusative case marker as well as the locative nominalizer *-an and the locative case marker removes any doubt of a connection between the case marking and nominalizing functions.

We can see how these accusative and oblique/locative case markers function in the Formosan languages, Thao, Bunun and Amis, below. In (53a), an actor voice predicate assigns accusative case to the 1SG object, necessitating the form *yakin* rather than the nominative *yaku*. In (53b), the predicate is now in an undergoer voice and therefore assigns nominative case to the patient. Under these circumstances, nominative *yaku* is required rather than accusative *yakin*.

Table 3. Comparative evidence for acc pronominals (Ross 2006: 533)

		1S	2S	1IP	1EP	2P
PAn	ACC	*i-ak-**ən**	*iSu[qu]-**n**	*ita-**ən**	*i-am-**ən**	*i-mu[qu]-**n**
Saisiyat	ACC	yak-**in**	ʔi-ʃoʔo-**n**	…	…	…
Thao	ACC	yak-**in**	ihu-**n**	ita-**n**	[y]am**in**	…
Paiwan	NOM	=[a]k**en**	=[e]su**n**	=[i]c**en**	=[a]m**en**	=[e]mu**n**
P-Bunun	ACC	*ðak-**un**	*suʔu-**n**	*it-**un**	*ðam-**un**	*muʔu-**n**
PMP	PSR	*[y]ak**ən**	…	*[y]at**ən**	*[y]am**ən**	…

13. Ross (2006: 535) considers the locative pronouns with *-an to have been innovated independently because the pronominal bases do not correspond with each other across languages. He infers from this that the PREP N-*an* construction existed in PAn but that the locative case marker had not yet morphologized on pronouns. Note that *-an appears to have been a case assigned by a variety of prepositions or predicates rather than an adposition itself.

Table 4. Preliminary reconstruction of PAn personal pronouns (Ross 2006: 532)

	1S	2S	3S	1IP	1EP	2P
NEUT	i-aku	iSu[qu]	s-ia	ita	i-ami	i-mu[qu], i-amu
NOM1	aku	Su[qu]	ia	(i)ta	ami	mu[qu], (amu)
NOM2	=ku, =[S]aku	=Su	(-ya)	=(i)ta	=mi[a], =[S]ami	=mu
ACC	**i-ak-ən**	**iSu[qu]-n**	**...**	**ita-ən**	**i-ami-n**	**i-mu[qu]-n**
GEN1	=[a]ku	=Su	(-ya)	=(i)ta	=mi[a]	=mu
GEN2	(=)m-aku	(=)m-iSu	...	(=)m-ita	((=)m-ami)	(=)m-amu
GEN3	n-aku	n-iSu	nia	ni-ta	ni-am, n-ami	ni-mu, n-amu

(53) Thao (Wang 2004: 232)

a. Ata kurkur yakin/*yaku!
NEG.IMP tickle.INTR.IMP 1SG.ACC/1SG.NOM
'Don't tickle me!'

b. Ata kurkur-i yaku/*yakin!
NEG.IMP tickle.TR.IMP 1SG.NOM/1SG.ACC
'Don't tickle me!'

We find a similar use of an accusative case marked pronoun in the Bunun example in (54) and the oblique case in Amis (55).

(54) Takituduh Bunun (Li 1997: 465)
ma-ludaq ʔasu ðakun
AV-hit 2SG.NOM 1SG.ACC
'You hit me.'

(55) Amis (Li 1997: 470)
mi-cahiw ci raraq ci tefiq-an
AV-persecute PM Rara PM Tefiq-OBL
'Raraq persecuted Tefiq.'

It seems that this productive use of noun phrase case marking with *-en and *-an was lost in Proto-Malayo-Polynesian. However, as Givón (2011: 97) notes, "Pronouns, including demonstrative pronouns, are notorious diachronic graveyards where relics of older case-marking morphology survive long after they have been lost in nouns (see e.g. English or German)" and this is certainly true for Austronesian languages, as well. Historically accusative pronouns form the base for

oblique pronouns in many Philippine languages so that the Tagalog example in (56) (glossed historically) is structurally cognate to the Formosan examples seen directly above.[14]

(56) Tagalog
T<um>ulong sila sa ak-in
<AV.PERF>help 3PL.NOM OBL 1SG-ACC
'They helped me.'

While this suffixal system of case marking is all but lost in Malayo-Polynesian languages, it remains a robust feature of Amis and several other Formosan languages. But putting aside its productivity in modern languages, the clear historical connection between these two undergoer nominalizers and objective case morphology demands explanation.

4.4 Towards an account of the case/nominalizer syncretism

Having established a case/nominalizer syncretism in Austronesian and beyond, we are now in a position to ask what can account for it. With regard to the pronominal paradigms reviewed above, Ross (2002: 46) puts forth the possibility that pronouns are of a fundamentally different semantic type in Austronesian than they are in Indo-European languages; rather than being individuals, they are predicates:

> The most striking piece of evidence that most content words were predications in PAn, or perhaps at some pre-PAn stage, was mentioned in Section 3.1: personal pronouns took the voice markers *-en (or *-n) and *-an (or *-nan). This suggests that, like the corresponding root in Straits Salish (Jelinek and Demers 1994: 715), second person singular *Su, for example, was a content word whose meaning might be translated as 'be you', *i-Su a phrase meaning 'the one who is you' (*i- being a determiner), *Su-n a content word meaning '[the one that] is you-ed', i.e. '[the one that] is yours', and *Su-(n)an a content word meaning '[the place that] you are at'.

14. What is not accounted for here is how the same set came to be used as independent genitive pronouns, which can function as agents of undergoer voice verbs, as in (i). In this case, our historical glossing, maintained for consistency, appears synchronically incoherent.

(i) Tagalog
Ak-ing t<in>ulung-an ang mga matanda.
1S-ACC:LNK <PERF>help-LV NOM PL old
'I helped the old person.'

This is a *prima facie* attractive idea, especially for the locative forms with *-an, since a very plausible pathway exists from locative to oblique/dative. Tibeto-Burman evidence points to *compounding* as a possible bridge between a word meaning 'place' (e.g., Tibetan *sa* and possibly Tani *kó*) and a locative nominalizer employed in relative clauses. Aldridge (2023) proposes a similar origin for PAn *-*an* as having grammaticalized from a noun meaning 'place' (and we can even point to a candidate here, PAn *zalan 'path', although this is obviously tentative without much hope for confirmation). Hendery (2012) furthermore shows how such words become general relativizers in a wide variety of languages, as shown in Table 5 (cf. Heine & Kuteva 2002).[15]

Table 5. Generic nouns as relative clause markers (from Hendery 2012: 63)

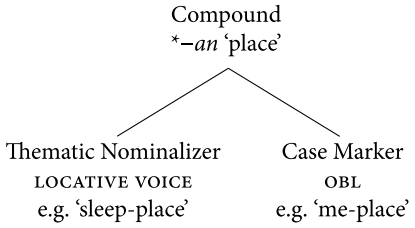
Language	Marker	Source	Details in
Thai	*thîi*	'place'	(Kuno 1981)
Lhasa Tibetan	*sa*	'earth, place'	(DeLancey 1986)
Malay	*tempat*	'place'	(Cumming 1991)
Hebrew	(*a*)*she*(*r*)	'place'	(Givón 1975; von Soden 1969; Berman 1997)
Japanese	(*mo*)*no*	'thing'	(Kuno 1981; Horie 1993)
Korean	*kes*	'thing'	(Simpson & Wu 2001)
Malay	*waktu*	'time'	(Cumming 1991)
Malay	*saat*	'moment'	(Cumming 1991)
Newari	*mha*	'body'	(DeLancey 1986)

The historical development shown schematically in (57), the left side of which was proposed earlier by Aldridge (2023), seems eminently plausible. It is, however, less obvious how to apply the compounding idea to patient nominalization/accusative case.[16]

15. Aldridge (2023) argues that the hypothetical 'place' noun gave rise to a general nominalizer directly without passing through a stage in which it was restricted to a locative function. However, the fact that it carries a distinctly locative function in many daughter languages, in addition to the many general nominalizer cases discussed by Aldridge (2023), would force an analysis in which a general nominalizer formed with 'place' became locative only after the 'place' etymology was completely obscured. While not impossible, this analysis does not appear to be the most parsimonious, especially in light of ample evidence for the change LOCATIVE > GENERAL NOMINALIZER.

16. Aldridge (2023) proposes that *-en patient voice is not reconstructable to PAn and has its origins in a daughter branch as a telicity marker. Aldridge does not consider the case marking functions of these suffixes but her historical analysis of both Austronesian *-en and *-an are relevant to the current discussion. For reasons of space, a full comparison with the present proposal will have to await further work.

(57)



The problem with *-en is that an accusative marked pronoun, for instance the reconstructed PAn *(i-)ak-en 1S.ACC form, cannot be described as '[the one that] is me-ed', as suggested by Ross, in structures such as (53) and (54), where it functions as an object.

A tantalizing connection to the Cuzco Quechua phenomenon termed "case float" by Lefebvre and Muysken (1988) allows us to bypass the difficulty in reconciling the form and interpretation of these syncretic markers. Recall from Section 3.2 that Quechua has a basic two-way distinction between subject and non-subject relativization marked by *-q* for subject relativization/nominalization and *-sqa/-na* for the realis and irrealis forms of non-subject relativization/nominalization (Lefebvre & Muysken 1988: 167). All varieties of Quechua have a basic SOV word order with nominative–accusative case marking, as shown in (58).

(58) Cuzco Quechua (Lefebvre & Muysken 1988: 18; Itier 1997: 66)
a. Xwan papa-ta mikhu-n
Juan potato-ACC eat-3
'Juan eats potatoes.'
b. T'anta-ta mikhu-ni
bread-ACC eat-1SG
'I eat the bread.'

In the Imbabura variety of Quechua shown in (59), we find an internally-headed relative clause in which the relativized NP retains its accusative case marking from the nominalized verb.

(59) Imbabura Quechua (Cole 1982: 49)
[wambra wagra-ta randi-shka] ali wagra-mi
boy cow-ACC buy-RLS.NMLZ good cow-FOC
'The cow which the boy bought is a good cow.'

Crucially, Lefebvre and Muysken (1988: 187) claim that relativization licenses "case float", in which "the relativized NP occurs within the relative clause but its Case appears on the nominalized verb." They relate this to the structure in (60a) where the notional head of the relative clause appears to move to a rightwards position while maintaining the expected accusative case marking from the relative clause,

without regard to the nominative case it might be expected to receive from the matrix clause. In (60b), the same NP, *warma* occurs in its supposed base position but now its case marker appears on the nominalized verb. This case float phenomenon can take place whether or not the relativized NP remains in its base position, as in (60b) or raises to a higher position at the edge of the relative clause, as in (60c).

(60) Cuzco Quechua (Lefebvre & Muysken 1988: 173)

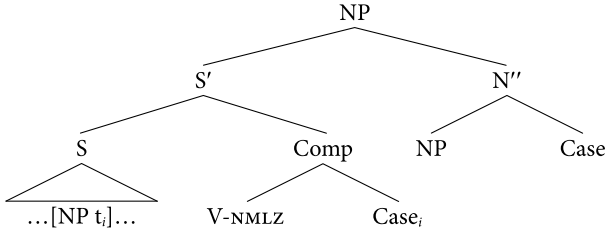
a. [Riku-sqa-y warma-ta] hamu-nqa.
see-NSBJ.RLS.NMLZ-1 girl-ACC come-3.FUT
'The girl I saw will come.'

b. [Warma riku-sqa-y-ta] hamu-nqa
girl see-NSBJ.RLS.NMLZ-1-ACC come-3.FUT

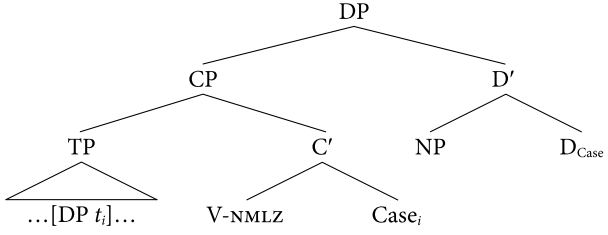
c. [Riku-sqa-y-ta warma] hamu-nqa
see-NSBJ.RLS.NMLZ-1-ACC girl come-3.FUT

Lefebvre and Muysken (1988) analyze the structure in (60b) as in (61), which we can update to the relabelled structure in (62).

(61)



(62)



The case marker raises from the relativized argument in the embedded clause to C, attaching to the nominalized verb. For the full motivation of this analysis, the reader is referred to the original work; what is important to us here is that the net result of case float is that a nominalized verb in a relative clause ends up marked with the case of the relativized NP. While case float provides a possible bridge from case marking to participant nominalization, there are two empirical problems here. The first is that the case marking on the nominalized verb is not the original participant nominalization itself (recall the subject nominalizer *-q* and non-subject nominalizers *-sqa/-na* from earlier). The second problem is even more basic: the key patterns in (60) in which the case marker attaches to the verb

have never been corroborated by other researchers despite several efforts (e.g. Hastings 2004).

Despite several open questions regarding Quechua case float, we have found unequivocal parallels to the Austronesian case/nominalizer syncretism in the Cahita languages and a strong basis for comparison in the Tani languages as well as Cuzco Quechua. This syncretism gives us a fresh motivation for considering the Ausrichtungsaffix or 'orientation affix' as a phenomenon quite separate from both grammatical voice and applicative morphology.

5. Equational Ausrichtungsaffix syntax vs. verbal applicative syntax

We began this investigation with a review of two paradigms reconstructed by Wolff (1973) and Ross (2002) for Proto-Austronesian, one set of which Starosta et al. (1982/2009) analyzed as deriving historically from participant nominalizations and the other as formed by suffixing applicatives that were grammaticalized from proclitic case markers or prepositions. Ross (2009) terms the original verbal forms the 'first generation' paradigm and the nominalization-derived forms the 'second generation' paradigm, because of their later adoption into matrix clauses as event-denoting predicates. He further identifies relative clauses as the key context for participle nominalizations and their expansion into matrix clauses as a key innovation characterizing a subgroup he names Proto-Nuclear-Austronesian. In the preceding, I have shown that participant nominalizations play an important role in the syntax of relative clauses (or their functional equivalents) cross-linguistically and that such an analysis has the power to explain other properties of conservative Austronesian languages, such as multiple determination (Section 3.4) and the case-marker syncretism (Section 4.4), which had not been previously considered relevant to this question. In this section, I argue that the syntax of first generation clauses versus second generation clauses must have differed considerably and bring new evidence from Philippine languages to bear.

We begin with Himmelmann (1987: 93–94), who makes a key connection between the nature of the Tagalog Ausrichtungsaffixe to the language's equational clause structure (emphasis mine):

> It should be noted that orientation in Tagalog occurs in a different morphosyntactic environment and serves different purposes than the formation of Indo-European verbal nouns. Thus, in Tagalog, it is not a 'peripheral' area of nominalization, but the central morphosyntactic procedure. Furthermore, one cannot start with a verb and its argument slots. Rather, the starting point is a conceptual word that denotes the state of affairs in a very general way, without specific orientation [...]. This general expression of the state of affairs can

> then be aligned with the help of orientation affixes so that it denotes one of the participants of the state of affairs. The aligned expression itself is neither referential nor necessarily predicative, but again a mere conceptual word that can be referentialized by 'ang' or attributed as a predicate to the referent (in the 'ang'-phrase). This alignment process goes hand in hand with the equational structure of the Tagalog sentence as described above. In such a sentence, a referentialized participant is attributed a state of affairs that is directed toward it and optionally further specified by attributes. In other words, alignment allows the complex relationships of a state of affairs with multiple participants, which usually require a verbal clause with a relational verbal center, to be expressed in the form of an equational clause, which is, in principle, tailored to the less complex relationships of a nominal clause (subject and predicate).

The equational clause structure goes hand in hand with the rich system of orientation morphology in the familiar way seen in (63). A bare participant nominalization, shown in (63a), becomes the predicate of a matrix clause, as in (63b) through an equational type structure.

(63) Tagalog
  a. ang=h<in>ukay-an
     NOM=<BEG>dig-LOC
     'the dug out area'
  b. h<in>ukay-an ni Totoy ang lupa.
     <BEG>dig-LOC PM.GEN Totoy NOM ground
     'Totoy dug out the ground'/'The ground was Totoy's digging place.'

This is parallel to the Paiwan example seen earlier in (1), used to illustrate the reanalysis of equational sentences with nominalized predicates as ordinary main clause assertions.

(64) Paiwan (Ferrell 1982; Ross 2002)
  a. təkəl-ən a vaua
     drink-PAT.NMLZ SPEC wine
     'the wine is something to be drunk'
  b. təkəl-ən nua qała a vaua
     drink-PAT.NMLZ GEN stranger SPEC wine
     'the wine is the stranger's drinking thing' →
     'the stranger will drink the wine'

The equational analysis of Philippine languages has a long pedigree and is clearly endorsed by Schachter and Otanes (1982: 62) in their Tagalog reference grammar. However, there exists a largely unexplored corollary of Himmelmann's insight regarding the tight connection between the Ausrichtungsaffixe and the equational clause structure which juxtaposes a predicate with a subject-like phrase (termed

'predication base' in Himmelmann (1987)). Namely, the licensing of a subject in any basic sentence such as (64) or (63) should not depend on any case assigning property of the predicate itself, as it might in English, or any other Indo-European language, but rather from an equational operator at the clause level. Whether this is treated as a null copula or something even simpler is immaterial to the point. But the first generation paradigm, which did not descend from nominalization and thus should lack the properties of the true Ausrichtungsaffixe, should differ in not showing the hallmarks of equational clause structure. In Philippine languages, the first generation paradigm lives on to varying degrees in imperatives, negated indicatives, consecutive clauses and constructions with aspectual auxiliaries. In the descriptive literature, the first and second generation paradigms are often treated tacitly as contextually determined allomorphs of each other, following Wolff (1973) (e.g., Cebuano *-an* LOCATIVE VOICE INDICATIVE vs. *-i* LOCATIVE VOICE IMPERATIVE). However, if these paradigms differed in lexical category rather than simply inflectional features, we expect deeper differences in the behavior of first and second generation forms, and it is to this evidence we now turn.

5.1 Imperative addressees

Kaufman (2009a,b) points out that second person singular addressees of first generation imperative verbs are obligatorily omitted in a range of Philippine languages but not omitted with second generation imperatives (Himmelmann 1999: 245), which often co-exist in the same language, as in southern dialects of Tagalog, as illustrated in (65).

(65) Batangas Tagalog
  a. Buks-i (*mo) ang pintuan
  open-LV.DEP 2SG.GEN NOM door
  'Open the door!'
  b. Buks-an *(mo) ang pintuan
  open-LV 2SG.GEN NOM door
  'Open the door!'

While seemingly a minor detail, this represents an important link between case licensing and Mood in first generation clauses, which does not generally exist in second generation (equational) clauses. If genitive case in second generation clauses is truly possessor case, we expect it to be governed by the nominalized predicate itself and not by higher clausal functions. If, on the other hand, it is a more typical verbal clause, we might expect that case assignment interacts more intimately with higher functions such as mood and tense that often tend to be associated with finiteness cross-linguistically.

5.2 Licensing nominative case in first generation clauses

McGinn (1988) notes a correlation between the presence of a nominative subject/pivot and a voice marked predicate in Tagalog. When there is voice marking on the predicate, the argument selected by voice must take the nominative case if it is overt. When the predicate happens to lack voice morphology, as in recent perfective clauses and gerunds, the clause also lacks the ability to assign nominative case (cf. Hsieh 2020). However, neither McGinn (1988) nor Hsieh (2020) note a systematic exception to this otherwise strong generalization which comes from first generation verbs. Here, we find that selection by the voice morphology does not, in fact, necessitate nominative case. We examine examples from four Philippine languages below.

5.2.1 *Bikol recent perfective*

Unlike the Tagalog recent perfective, which lacks any hint of a voice distinction and assigns genitive case to all arguments, the cognate construction in certain related languages makes use of the first generation paradigm to indicate diathesis. Crucially, however, no nominative case is assigned in these clauses, as seen in (66), from Miraya Bikol.

(66) Miraya Bikol (Stevens 1969: 13)

a. 1ST GENERATION ACTOR VOICE (*-Ø*)
ka-ʔiː-ʔinum=ko=la han túbig
RCNT-RCNT-drink=1SG.GEN=only OBL water
'I just drank water.'

b. 1ST GENERATION LOCATIVE VOICE (*-i*)
ka-ʔiː-ʔinum-i=pa=sana=nya kani=ŋ buːti=ŋ ini
RCNT-RCNT-drink-LV.DEP=still=only=3SG.GEN OBL=LNK bottle=LNK this
'He just drank from this bottle.'

c. 1ST GENERATION PATIENT VOICE (*-a*)
ka-bá-bakal-a=ku=pa=sana ka-ini sa-iya
RCNT-RCNT-buy-PV.DEP=1SG.GEN=still=only OBL-this OBL-3SG
'I just bought this from him.'

In (66b) and (66c), the first generation morphology selects the locative and patient arguments, respectively, yet they appear in the oblique case, unlike what we would find in a regular clause with a predicate marked by second generation morphology. These types of recent perfective constructions in Philippine languages are known to lack properties of normal finite predications and resemble non-finite exclamatives. For instance, unlike normal predications, they cannot be negated. We return to this point below.

5.2.2 *Waray temporal adjuncts*

Similarly, in Waray temporal adjuncts, we find the familiar productive diathesis alternations of Philippine-type languages but without assignment of nominative case. In (67a), *baŋku* 'bench' is selected as the pivot by the locative voice, but it is assigned genitive case. In (67b), *iya kamot* 'his hand' is selected by the patient voice but also assigned genitive case instead of nominative.

(67) Waray (Zorc 1977: 139; Society 1984)

a. 1ST GENERATION LOCATIVE VOICE (*-i*)
pag-liŋkur-i=niya han baŋku, na-ruba?
SBJNC-sit-LV.DEP=3SG.GEN GEN.DEF bench STAT.PERF-break
'When he sat on the bench, it broke.'

b. 1ST GENERATION PATIENT VOICE (*-a*)
Pero pag-bunlot-a=gud=niya tipabalik han iya
but SBJNC-pull-PV.DEP=EMPH=3SG.GEN back GEN.DEF 3SG.GEN
kamot…
hand
'But as soon as he drew back his hand…' (Genesis 38:29)

5.2.3 *Western Bukidnon Manobo*

Western Bukidnon Manobo, as described by Elkins (1970), employs second generation predicates for regular main clause assertions with the familiar Philippine-type diathesis seen in (68) but uses first generation predicates in a number of contexts including two different types of nominalizations. In (69), we see one a gerundial type nominalization where the predicate still displays a voice alternation, signaled by zero in the actor voice, *-a* in both patient voice and conveyance voice, and *-i* for the locative voice. Despite the active diathesis alternation, all arguments are uniformly marked with the oblique case in this construction.

(68) Western Bukidnon Manobo (Elkins 1970: 38)

a. 2ND GENERATION ACTOR VOICE (Ø)
eD-badti'=a te rasang
IMPF-[AV]break=1SG.NOM OBL twig
'I break the twig.'

b. 2ND GENERATION PATIENT VOICE (*-en*)
eD-badti-en=ku is rasang
IMPF-break-PV=1SG.GEN NOM twig
'I break the twig.'

c. 2ND GENERATION CONVEYANCE VOICE (*i-*)
i-D-badti=ku is sipit
CV-IMPF-break=1SG.GEN NOM pliers
'I use the pliers to break (it).'

d. 2ND GENERATION LOCATIVE VOICE (*-an*)
eD-badti-an=ku sikew
IMPF-break-LV=1SG.GEN 2SG.NOM
'I break (it) for you.'

(69) Western Bukidnon Manobo (Elkins 1970: 38)
a. 1ST GENERATION ACTOR VOICE (Ø)
k-eD-badti'=ku te rasang
NMLZ-IMPF-[AV.DEP]break=1SG.GEN OBL twig
'my breaking the twig'
b. 1ST GENERATION PATIENT VOICE (*-a*)
k-eD-badti-a=ku te rasang
NMLZ-IMPF-break-PV.DEP=1SG.GEN OBL twig
'my breaking the twig'
c. 1ST GENERATION CONVEYANCE VOICE (*-a*)
k-eD-badti-a=ku te sipit
NMLZ-IMPF-break-CV.DEP=1SG.GEN OBL pliers
'my using the pliers to break'
d. 1ST GENERATION LOCATIVE VOICE (*-i*)
k-eD-badti-i=ku kenikew
NMLZ-IMPF-break-LV.DEP=1SG.GEN 2SG.OBL
'my breaking for you'

We see a similar contrast between the main clause second generation predicates in (70) and the first generation predicates in (71), used here in a nominalization indicating the reason for an action. As in (69), none of the arguments in (71) receive nominative case despite being selected by the verbal morphology.

(70) Western Bukidnon Manobo (Elkins 1970: 40)
a. 2ND GENERATION ACTOR VOICE (Ø)
eD-'uwit=a kenikew
IMPF-[AV]carry=1SG.NOM 2SG.OBL
'I carry you.'
b. 2ND GENERATION PATIENT VOICE (*-en*)
eD-'uwit-en=ku sikew
IMPF-carry-PV=1SG.GEN 2SG.NOM
'I carry you.'
c. 2ND GENERATION LOCATIVE VOICE (*-an*)
eD-'uwit-an=ku sikew te vegas
IMPF-CARRY-LV=1SG.GEN 2SG.NOM OBL rice
'I carry some rice to you.'

(71) Western Bukidnon Manobo (Elkins 1970: 40)

a. 1ST GENERATION ACTOR VOICE (Ø)
i-D-'uwit=ku kenikew
NMLZ-IMPF-[AV.DEP]carry=1SG.GEN 2SG.OBL
'my reason for carrying you'

b. 1ST GENERATION PATIENT VOICE (*-a*)
i-D-'uwit-a=ku kenikew
NMLZ-IMPF-carry-PV.DEP=1SG.GEN 2SG.OBL
'my reason for carrying you'

c. 1ST GENERATION LOCATIVE VOICE (*-i*)
i-D-'uwit-i=ku kenikew te vegas
NMLZ-IMPF-carry-LV.DEP=1SG.GEN 2SG.OBL OBL rice
'my reason for carrying rice to you'

5.2.4 *Bugkalot*

Finally, in Bugkalot (also known as Ilongot or Ėgongot), a South Cordilleran language of northern Luzon, we find the first generation paradigm employed in consecutive clauses, imperatives, as well as an apparent auxiliary construction using the light verb *aġa* 'take'.[17] In imperatives and the auxiliary construction, the first generation forms assign nominative case to the argument they select, as expected. (Note that Bugkalot does not preserve the historical ban on expressing second person singular agents in imperatives with first generation forms discussed in Section 5.1.) Specifically, *abong* 'house' in (72a) receives nominative case thanks to the patient voice signaled on the verb by *-a*. The first person pronoun in (72b) receives nominative case thanks to the locative voice *-i*.

(72) Bugkalot

a. 1ST GENERATION PATIENT VOICE (*-a*)
in-aġa=tu tunuġ-a ima abong
PERF-take=3SG.GEN burn-PV.DEP NOM house
'He did burn the house.'

b. 1ST GENERATION LOCATIVE VOICE (*-i*)
bëgë'-i=mo=ak
say-LV.DEP=2SG.GEN=1SG.NOM
'Tell me!'

However, in consecutive clauses (indicated by a suffix *-i* that appears unrelated to the locative voice *-i*), we find first generation morphology without nominative case on the selected argument. This is exemplified in (73), where the locative

17. The Bugkalot examples cited here are thanks to Frederick Barcelo of Pëlaway and were recorded by the author in July of 2025 in Bambang.

voice selects *Madiya* 'Maria' as the recipient, although this argument can only be expressed in oblique case rather than the nominative.[18]

(73) Bugkalot
bëgë-i-yi=to nod Madiya
say-LV.DEP-PURP=3S.GEN PM.OBL Maria
'so that he says (it) to Maria.'

5.3 Two means of assigning case

The generalization to be made over the examples reviewed in Section 5.1 and Section 5.2 is that, while we find similar diathesis alternations across the first and second generation paradigms, they appear to make use of different case assignment mechanisms. In second generation clauses, genitive and oblique case appears to be assigned by the predicate itself without any potential interference from higher functional categories. The nominative case associated with the pivot argument, on the other hand, is assigned by a (covert) equational operator which has no relation to tense, mood or speech act. It does, however, require that the two phrases be in a particular syntactic relation to each other, as would be expected. The predicate must be in a dedicated clause-initial position while the pivot with which it is equated takes a position following the predicate and the genitive agent in Philippine languages generally. On the other hand, nominative assignment by a first generation predicate requires more than a particular syntactic configuration; it depends on high-level features of the clause related to mood and finiteness. In the recent perfective, consecutive clauses, temporal adjuncts, and nominalizations, we find first generation voice alternations without concomitant nominative case on the pivot because this additional dependency fails to be met.

The structural requirements on case assignment in second generation forms can also fail to be met. Recall that the equational structure requires the second

18. A caveat is necessary here regarding common processes of conflation and loss of the first generation paradigm. In Bugkalot, for instance, we do not find the expected combination of the first generation patient voice suffix *-a with the innovative consecutive suffix *-i*. Where we expect a reflex of *-a we again find *-i*, the first generation locative suffix representing the patient voice. (Note, however, that there is a reconstructed shift from patient voice to locative voice in negated and possibly other irrealis contexts in early Austronesian, cf. Ross 2009, 2008.) We also find combinations of consecutive *-i* with the second generation conveyance *i-* prefix as well as actor voice *m-*. In the first case, this is because Bugkalot, like the vast majority of Philippine languages, has lost the first generation conveyance suffix *-an. In the second case, actor voice *<um> and its allomorphs have an ambiguous position with regard to their membership in the first and second generation paradigms and seem to have existed in both.

generation form to be in clause-initial position, generally referred to as the predicate position. However, a range of auxiliaries, negators, adverbs and interrogatives may also occupy this position. In these cases, provided that the language maintains both paradigms, we often find that the first generation forms are required. In the Cebuano examples in (74), the predicate position (indicated by underlining) is occupied by a negator in (a), an aspectual/temporal predicate in (b) and (c), and an interrogative in (d). In all of these circumstances, second generation forms are blocked and first generation forms are required. The same situation is shown in (75), from another Philippine subgroup.

(74) Cebuano (Wolff 1965: 187,189,193)

a. Wa' taga:'-i ang ba:ta'
NEG.EXT give-DEP.LV NOM child
'The child was not given any.'

b. Human na na' na:ku' lutu:'-a
finish IAM that 1SG.GEN cook-DEP.PV
'I have finished cooking that.'

c. Kaduha ku tawg-a si Huwan
twice 1SG.GEN call-DEP.PV PM.NOM Juan
'I called John twice.'

d. di'in ni:mu kuha:'-a kana=ng i:mu=ng kuwarta?
where.PST 2SG.GEN get-DEP.PV that=LNK 2SG.GEN=LNK money
'Where did you get that money of yours from?'

(75) Western Bukidnon Manobo (Elkins 1970: 54)
diye' tew eD-pumpun-a is menge etew
there 1PL.IN.GEN IMPF-collect-DEP.PV NOM PL person
'We'll all go collect the people there.'

I interpret these facts as showing that the equational structure associated with the second generation forms is crucially reliant on these forms occupying the predicate position. The first generation forms have no such requirement and thus appear when this position is otherwise occupied.[19] This explains why such an otherwise arbitrary set of elements (negation, focused phrases, auxiliaries) require the first generation forms in Philippine languages that preserve both paradigms; the trigger is primarily positional rather than semantic.

19. It is due to this property that the first generation forms are termed the dependent paradigm by Wolff (1973), as they typically occur in the presence of a preverbal functor.

5.4 Genitive case in first generation clauses

It is somewhat of an embarrassment to Starosta et al.'s (1982/2009) hypothesis that agents of first generation clauses also receive genitive case despite never having passed through a stage of nominalization (Jiang 2016: 201), as nominalization is ostensively the basis for the widespread identical treatment of non-actor voice agents and possessors in Philippine-type languages. Two points must be made here. First, it is remarkable that the Formosan languages which still employ the first generation paradigm in declarative main clauses (Puyuma, Rukai and Tsou) do not employ the *n- initial genitive case formant generally found in languages that have adopted the nominalized second generation paradigm as main clause predicates (i.e., Ross's Nuclear Austronesian subgroup). Instead, they use formants associated with oblique relations (*k- and *t-, see Ross 2006) despite also applying these markers to possessors. In other words, while nearly all Formosan languages show a syncretism between possessors and non-actor voice agents, only those languages that have reanalyzed nominalizations as regular main clause predicates do so with the historical possessor case. This may have resulted from paradigmatic leveling across the two paradigms, an idea which finds support from Kanakanavu, as discussed by Cheng (2024).

Tsuchida (1976); Wild (2018); Zeitoun and Teng (2016); Teng and Zeitoun (2016); Cheng (2023, 2024) note that the Kanakanavu patient voice perfective *-ai*, a first generation form, expresses its agent in the oblique case rather than the genitive, as shown in (76).

(76) Kanakanavu (Cheng 2024: 221)
pana'-ai 'ikua
shoot-PV.PERF 1SG.OBL
'I shot (it).'

Cheng (2024) argues convincingly that this case assignment should be seen as a retention rather than an innovation and thus represents a missing link between the original case assignment pattern of first generation clauses and the agent-possessor syncretism they widely show in modern languages. Specifically, agents of first generation non-actor voice verbs were originally expressed in the oblique case but merged with the genitive in most languages due to analogical pressure from second generation clauses, which took on an increasingly large share of the syntax. For Kanakanavu to have shifted in the other direction, from genitive to oblique, would seem entirely unmotivated.

5.5 First generation as actor voice applicatives in Malayo-Polynesian languages

In the overall picture, the distinction between Philippine-type voice and applicatives is not always clear (see Riesberg (2014); Himmelmann and Riesberg (2013); Kaufman (2017) for discussion). For present purposes, we only consider unambiguous cases of applicatives as those that co-occur in an actor voice clause on a predicate that is simultaneously marked by actor voice morphology (i.e., a descendant of PAn *<um>), a phenomenon that is typical of western Malayo-Polynesian languages south of the Philippine zone.[20] In many Malayo-Polynesian languages beyond the Philippines, first generation suffixes become bona fide applicatives in the aforementioned sense, co-occurring with actor voice. For locatives, recall that the first generation suffix is *-i while the second generation suffix is *-an. In a nearly atavistic return to Starosta et al.'s (1982/2009) proposed origins of *-i as a captured preposition introducing a following locative phrase, Indonesian/Malay *-i* functions to promote oblique arguments and adjuncts to object position regardless of voice. Meanwhile, the Indonesian/Malay suffix *-an*, derived from the second generation suffix *-an, also returns to its origins to again function as a general nominalizer outside of the system of verbal predication, as seen in (77). Combining actor voice verbs with second generation morphology, as in (77c), is impossible in the vast majority of languages that allow for true applicatives.[21]

(77) Indonesian
  a. men-jalan-i
     AV-walk-LOC.APPL
     'to walk on/to'
  b. jalan-an
     walk-NMLZ
     'road'
  c. *men-jalan-an
     AV-walk-NMLZ

20. Zobel (2002) posits this "Janus-faced" voice marking as a defining feature of a large subgroup spanning from western Indonesia to Chamorro in the east. A different type of combination is unique to the languages of northern Luzon in the Philippines, whereby a second generation conveyance voice prefix *i-* combines with the actor voice. This scope and syntax of this applicativization in Philippine languages is in much need of further study but cannot be considered here.

21. The *-an* suffix which is found with actor voice predicates, as in Totoli descends from the first generation conveyance voice (Himmelmann and Riesberg 2013: 402, Footnote 8, where it is referred to as the "nonindicative (or atemporal) circumstantial (instrumental) voice suffix").

These developments are far from unique to Indonesian/Malay and span a large number of subgroups throughout the Indo-Malaysian archipelago (see fn. 22). This split is once again predicted if second generation forms maintain the original equational syntax in their predicate function while first generation forms continue a more canonical verbal predication in which argument structure can be altered via cross-linguistically familiar morphosyntactic operations (e.g., applicativization, passivization, etc.).

6. Conclusion and remaining questions

In the spirit of Himmelmann (1987, 1991), I have attempted to establish the Ausrichtungsaffix or orientation affix as a distinct type of diathesis which has roots in participant nominalization. In addition to displaying a cluster of unique properties described in Section 3, it also appears to bridge voice and case across several language families, including Austronesian, meriting its own place in grammatical ontology as distinct from voice, applicative marking and agreement. Importantly, the richness of Philippine-type voice systems, with four basic alternations, appears to find analogues in other languages that have recruited participant nominalizations for use in relative clauses. At the same time, Austronesian uniquely marries two types of diathesis in a single grammatical system: the so-called first generation paradigm, which does appear to have roots in applicatives, as first posited by Starosta et al. (1982/2009), and the second-generation forms, which has the aforementioned nominal properties.

Many of the facts laid out here represent a standing challenge to those analyses that treat diathesis as agreement as well as those that treat the distinction between first and second generation forms as shallow allomorphy. However, there are questions posed by Jiang (2016) and Chen (2017) in advocating this view that cannot be easily brushed away. One problem involves the occurrence of first generation forms in relative clauses in languages that generally require second generation forms for this purpose. Jiang (2016) notes that Rikavung Puyuma, like other Puyuma dialects, uses second generation forms for relative clauses in the unmarked case but switches to first generation forms in the presence of negation, as shown in (78).

(78) Rikavung Puyuma (Jiang 2016: 189, glosses adapted for consistency)
a. mau mau i=misak [na='azi=ku kezeng-i]
EMPH COP NOM.PM.SG=M. NOM.DEF=NEG=1SG.ACT pull-LV.DEP
'The (only) one I didn't pull is Misak.'

b. mau mau i=misak [na='azi=ku sukun-an]
EMPH COP NOM.PM.SG=M. NOM.DEF=NEG=1SG.ACT push-CV.DEP
'The (only) one I didn't push is Misak.'

He shows the same phenomenon for Seediq and Siraya, two other Formosan languages, in (79) and (80), where the first generation form is not triggered by negation but by the irrealis mood of the relative clause.

(79) Truku Seediq (Tsukida 2009: 236, cited by Jiang 2016: 204)
yayu niyi ka=[keret-ani sagas]
knife PROX TOP=cut-CV.DEP watermelon
'What one should cut the watermelon with is this knife.'

(80) Siraya (Adelaar 2011: 91, cited by Jiang 2016: 226)
mang ta [kan-aw=miän]? mang ta [ĭt-aw=miän]?
what TOP eat-PV.IRR=1PL.EXCL.ACT what TOP drink-PV.IRR=1PL.EXCL.ACT
'What are we going to eat? What are we going to drink?'

To add a parallel example from Central Philippine languages, we can see how an imperative embedded in a noun phrase maintains first generation morphology rather than second generation morphology, although, as seen earlier in (65), a configuration such as (81b) is ungrammatical in related languages.

(81) Tausug (Akamine 2003)
a. Kaun-a ini!
eat-PV.DEP this.NOM
'Eat this!'
b. Ini in kaun-a!
this.NOM NOM eat-PV.DEP
'This is what you eat!'

Similarly, in (82), the verb takes the first generation locative *-i* suffix instead of *-an* due to the presence of negation, despite being embedded in a determiner phrase.

(82) Cebuano
ang=wala nako ma-uyon-i…
NOM=NEG 1SG.GEN PV.POT-endorse-LV.DEP
'What I can't endorse…'

At the same time, the nominal context does not obviate other properties of first generation forms. For instance, the imperative addressee is still obligatorily omitted in a language like Tausug, even when the imperative is embedded in a determiner phrase, as seen in (83), showing that the encasing determiner phrase does not overwrite the syntactic properties of the verb.

(83) Tausug
*Ini in kaun-a mu!
this.NOM NOM eat-PV.DEP 2SG.GEN
(For, 'This is what you eat!')

There are several ways to approach these facts, none of which I can explore satisfactorily here. The simplest is to say that the strong restrictions on verbal subordination we find in certain Formosan languages can be overridden when a higher clausal function (relating, for instance, to speech act) needs to be expressed within an argument phrase. On this analysis, the exceptions seen above actually contain a larger clausal constituent than is usually found with plain second generation relatives, which should be a testable hypothesis. But regardless of how the Puyuma data is analyzed, it is clear that the first vs. second generation distinction has been reinterpreted in a number of ways across Austronesian languages. In many languages, it does appear to have become a realis vs. irrealis distinction, with the former indicated by second generation and the latter indicated by first generation morphology. The difference between the position staked out here and that of Jiang (2016), Chen (2017) and Blust and Chen (2017) is that I believe mood distinctions to have been wholly derived from what was originally a purely syntactic distinction: when the predicate position was otherwise occupied, it would be impossible for the second generation form to constitute an equational structure, thus forcing use of the first generation paradigm. On the other hand, they believe the mood distinctions to be the original factor distinguishing these two paradigms, with apparent categorial effects a mere epiphenomenon. This invites a larger research program of examining differences between these two paradigms in living languages to see whether we can detect positional triggers, as commonly found in the Central Philippine languages, and other syntactic facts that bear no relation to mood or whether these apparent categorial differences can sensibly be reduced to mood based on broader typological evidence.

Acknowledgements

I am truly honored to be able to dedicate this chapter to Nikolaus Himmelmann, who has profoundly shaped my thinking from my earliest days as a student of Austronesian languages in the Philippines (where we discussed the work of 'Pareng Himmelmann' in reverential tones) to the present day. He has served as an inspiration to me not only in all matters Austronesian but also in the realms of language documentation, where he has had a transformative influence. I cannot thank him enough for his support, encouragement and collaboration over many years.

In addition to the editors of this volume and an anonymous reviewer, who helped improve the manuscript in many ways, I would also like to thank Edith Aldridge, Stacy Teng, Yi-Yang Chen, Paul Li, Mark Post and audiences at the UCSB Center for Taiwan Studies and the Fifth

World Congress of Taiwan Studies, where portions of this work were originally presented. Needless to say, none of the above necessarily share the views expressed here and should not be held responsible for any errors herein.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following abbreviations are used:

ACT	actor case	IN	inclusive
AGT	agent	INCP	incipient aspect
AV	actor voice	INSTR	instrumental
BEG	begun/initiated aspect	ITR	iterative
CAUS.INDIR	indirect causative	LNK	linker
CONV	conveyance	NSBJ	non-subject
CORE	core argument	NZP	nominalizing prefix
COS.ASP	change of state aspect	PAT	patient
CV	conveyance voice	PERF	perfective
DEP	dependent voice paradigm	PM	personal marker
DES	desiderative	POT	potentive
DEF	definite	PRT	participle
DEP	dependent paradigm (applicative-derived Austronesian diathesis markers)	PV	patient voice
		QM	question marker
		RCNT	recent perfective aspect
		RED	reduplication
DES	desiderative	REM	remote (aspect)
DIR	directional	RLS	realis
EMPH	emphatic	SBJNC	subjunctive
EXT	existential	SPEC	specific
IAM	iamitive (e.g., 'already')	STAT	stative
INDF	indefinite	UV	undergoer voice
IMPF	imperfective aspect	VT	transitive verb

References

Abraham, P. T. 1985. *Apatani Grammar*. Mysore: Central Institute of Indian Languages.

Ackerman, Farrell, and Irina Nikolaeva. 2013. *Descriptive Typology and Linguistic Theory: A Study in the Morphosyntax of Relative Clauses*. Stanford: CSLI Publications.

Adelaar, Alexander K. 2011. *Siraya: Retrieving the Phonology, Grammar and Lexicon of a Dormant Formosan Language*. Berlin: Walter de Gruyter.

Akamine, Jun. 2003. *Tausug Sentence List*. Unpublished ms.

Aldridge, Edith. 2004. Ergativity and Word Order in Austronesian Languages. Doctoral Dissertation, Cornell University.

Aldridge, Edith. 2023. "Diachronic Development of the Nuclear Austronesian Locative and Patient Applicatives." In *Applicatives in Austronesian Languages*, ed. by Jocelyn Aznar, Christian Döhler, and Jozina Vander Klok, 5–39. Tokyo: Tokyo University of Foreign Studies.

Álvarez González, Albert. 2012. "Relative Clauses and Nominalizations in Yaqui." In *Relative Clauses in Languages of the Americas: A Typological Overview*, ed. by Bernard Comrie, and Zarina Estrada-Fernández, 67–96. Amsterdam/Philadelphia: John Benjamins.

Álvarez González, Albert. 2016. "The Evolution of Grammatical Nominalizations in Cahita Languages." In *Finiteness and Nominalization*, ed. by Claudine Chamoreau, and Zarina Estrada-Fernández, 107–140. Amsterdam/Philadelphia: John Benjamins.

Álvarez González, Albert. 2019. "The 'Relative' Illusion and the Origin of Non-Subject Nominalizers in Cahita (Uto-Aztecan)." In *Nominalization in Languages of the Americas*, ed. by Roberto Zariquiey, Masayoshi Shibatani, and David W. Fleck, 301–340. Amsterdam/Philadelphia: John Benjamins.

Álvarez González, Albert. 2023. "The Origin of Instrumental Postpositions in Cahita (Uto-Aztecan): A New Account." *Cuadernos de Lingüística de El Colegio de México* 10: 1–50.

Horie, Kaoru. 1993. "Internally Headed Relative Clauses in Korean and Japanese: Where Do the Differences Come from?" In *Harvard Studies in Korean Linguistics* 5, ed. by Susumu Kuno et al., 449–458. Seoul: Hanshin Publishers.

Kuno, Susumu, and Preya Wongkhomthong. 1981. "Relative Clauses in Thai." *Studies in Language* 5 (2): 195–226.

Blust, Robert, and Victoria Chen. 2017. "The Pitfalls of Negative Evidence: 'Nuclear Austronesian', 'Ergative Austronesian', and their Progeny." *Language and Linguistics* 18 (4): 577–621.

Berman, Ruth. 1997. Modern Hebrew. In *The Semitic Languages*, ed. by Robert Hetzron, 312–334. London: Routledge.

Cumming, Susanna. 1991. *Functional Change: The Case of Malay Constituent Order*. Berlin: De Gruyter Mouton.

Givón, Talmy. 1975. On the Role of Perceptual Clues in Hebrew Relativization. Malibu: Undena Publications.

Capell, Arthur. 1964. "Verbal Systems in Philippine Languages." *Philippine Journal of Science* 93: 231–249.

Chen, Victoria. 2015. "A Reexamination of Recent Austronesian Higher-Level Subgroupings and the Nominalization-into-Verb Hypothesis." Presentation made at ICAL 13, Academia Sinica, July 22 2015.

Chen, Victoria. 2017. A Reexamination of the Philippine-Type Voice System and its Implications for Austronesian Primary-Level Subgrouping. Doctoral Dissertation, University of Hawai'i at Mānoa. http://hdl.handle.net/10125/62502

Chen, Victoria, and Shin Fukuda. 2017. "Relabelling 'Ergative': Evidence from Formosan." In *Proceedings of AFLA 23*, ed. by Nomoto, Hiroki, Takuya Miyauchi, and Asako Shiohara, 50–64. Canberra: Asia-Pacific Linguistics. http://hdl.handle.net/1885/111479

Cheng, Yi-Yang. 2023. "Tense-Aspect Conditioned Agent Marking in Kanakanavu, an Austronesian Language of Taiwan." *Open Linguistics* 9: 1–26.

Cheng, Yi-Yang. 2024. "On the Rise of an Unusual Austronesian Voice System: Perfective-Like Verb Forms in Kanakanavu." *Italian Journal of Linguistics* 36 (2): 213–240.

Chung, Sandra. 1994. "Wh-Agreement and 'Referentiality' in Chamorro." *Linguistic Inquiry* 25: 1–44. https://www.jstor.org/stable/4178847

Simpson, Andrew, and Zoe Wu. 2001. "The grammaticalization of formal nouns and nominalizers in Chinese, Japanese and Korean." In *Language Change in East Asia*, ed. by Thomas E. McAuley, 250–283. London: Curzon.

Chung, Sandra. 1998. *The Design of Agreement: Evidence from Chamorro*. Chicago: University of Chicago Press.

Cole, Peter. 1982. *Imbabura Quechua*. Amsterdam: North-Holland Publishing Company.

Comrie, Bernard. 1981. *Language Universals and Linguistic Typology: Syntax and Morphology*. Chicago: University of Chicago Press.

Comrie, Bernard. 1998. "Rethinking the Typology of Relative Clauses." *Language Design* 1: 59–86.

Comrie, Bernard, and Zarina Estrada-Fernández (eds). 2012. *Relative Clauses in Languages of the Americas: A Typological Overview*. Amsterdam/Philadelphia: John Benjamins.

Comrie, Bernard, and Sandra A. Thompson. 1985. "Lexical Nominalization." In *Language Typology and Syntactic Description*, vol. 3, ed. by Timothy Shopen, 349–398. Cambridge: Cambridge University Press.

Coupe, Alexander Robertson. 2007. *A Grammar of Mongsen Ao*. Berlin: Mouton de Gruyter.

Cristofaro, Sonia. 2019. "Nominalization in Cross-Linguistic Diachronic Perspective." In *Nominalization in Languages of the Americas*, ed. by Roberto Zariquiey, Masayoshi Shibatani, and David W. Fleck, 169–194. Amsterdam/Philadelphia: John Benjamins.

DeLancey, Scott. 1986. *Relativization as Nominalization in Tibetan and Newari*. Ms. 19th ICSTLL.

DeLancey, Scott. 1999. "Relativization in Tibetan." In *Topics in Nepalese Linguistics*, ed. by Yogendra P. Yadava, and Warren W. Glover, 231–249. Royal Nepal Academy.

Dryer, Matthew S. 1986. "Primary Objects, Secondary Objects, and Antidative." *Language* 62 (4): 808–845.

Elkins, Richard E. 1970. *Major Grammatical Patterns of Western Bukidnon Manobo*. Norman, Oklahoma: SIL.

Erlewine, Michael Yoshitaka. 2019. "Patterns of Relativization in Austronesian and Tibetan." Presented at Comparative Syntax in East and Southeast Asia, Sogang University, December 2019.

Erlewine, Michael Yoshitaka, Theodore Levin, and Coppe van Urk. 2017. "Ergativity and Austronesian-Type Voice Systems." In *The Oxford Handbook of Ergativity*, ed. by Lisa Travis, Jessica Coon, and Diane Massam. Oxford: Oxford University Press.

Estrada-Fernández, Zarina. 2008. "Cláusulas relativas em pima bajo." *LIAMES: Línguas Indígenas Americanas* 8 (1): 69–83.

Estrada-Fernández, Zarina. 2012. "From Demonstrative to Relative Marker to Clause Linker: Relative Clause Formation in Pima Bajo." In *Relative Clauses in Languages of the Americas: A Typological Overview*, ed. by Bernard Comrie, and Zarina Estrada-Fernández, 127–146. Amsterdam/Philadelphia: John Benjamins.

Evans, Nicholas, and Honoré Watanabe (eds). 2016. *Insubordination*. Amsterdam: John Benjamins.

Ferrell, Raleigh. 1982. *Paiwan Dictionary*. Canberra: Pacific Linguistics.

Foley, William A. 1998. "Symmetrical Voice Systems and Precategoriality in Philippine Languages." Paper presented at the 3rd LFG conference, Brisbane.

Foley, William A. 2008. "The Place of Philippine Languages in a Typology of Voice Systems." In *Voice and Grammatical Relation in Austronesian Languages*, ed. by Peter K. Austin, and Simon Musgrave, 22–44. CSLI Publications.

Genetti, Carol. 2007. *A Grammar of Dolakha Newar*. Berlin/New York: Mouton de Gruyter.

Gijn van, Rik. 2024. "Syntactic Structures in the Languages of the Central Andes." In *Oxford Guide to the Languages of the Central Andes*, ed. by Mattias Urban, 597–634. Oxford: Oxford University Press.

Gildea, Spike. 1998. *On Reconstructing Grammar: Comparative Cariban Morphosyntax*. Oxford: Oxford University Press.

Givón, Talmy. 2001. *Syntax: An Introduction*. Amsterdam: John Benjamins.

Givón, Talmy. 2011. *Ute Reference Grammar*. Amsterdam/Philadelphia: John Benjamins.

Givón, Talmy. 2012. "Toward a Diachronic Typology of Relative Clause." In *Relative Clauses in Languages of the Americas: A Typological Overview*, ed. by Bernard Comrie, and Zarina Estrada-Fernández, 3–26. Amsterdam/Philadelphia: John Benjamins.

Hankamer, Jorge, and Laura Knecht. 1976. "The Role of the Subject/Non-Subject Distinction in Determining the Choice of Relative Clause Participle in Turkish." *Harvard Studies in Syntax and Semantics* 2: 197–219.

Hastings, Rachel. 2004. The Syntax and Semantics of Relativization and Quantification: The Case of Quechua. Doctoral Dissertation, Cornell University, Ithaca, NY.

Haude, Katharina. 2025. "An Austronesian-Type Voice System in an Amazonian Isolate? Comparing Movima and Tagalog." In *Investigating Language Isolates: Typological and Diachronic Perspectives*, ed. by Iker Salaberri, Dorota Krajewska, Ekaitz Santazilia, and Eneko Zuloaga, 71–102. Amsterdam/Philadelphia: John Benjamins.

Heine, Bernd, and Tania Kuteva. 2002. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Hendery, Rachel. 2012. *Relative Clauses in Time and Space: A Case Study in the Methods of Diachronic Typology*. Amsterdam/Philadelphia: John Benjamins.

Himmelmann, Nikolaus P. 1987. *Morphosyntax und Morphologie. Die Ausrichtungsaffixe im Tagalog*. München: Fink.

Himmelmann, Nikolaus P. 1991. "The Philippine Challenge to Universal Grammar." *Institute for Linguistics, University of Köln Working Papers* 15: 1–60.

Himmelmann, Nikolaus P. 1999. "The Lack of Zero Anaphora and Incipient Person Marking in Tagalog." *Oceanic Linguistics* 38 (2): 231–269.

Himmelmann, Nikolaus P. 2005. "Typological Characteristics." In *The Austronesian Languages of Asia and Madagascar*, ed. by Alexander Adelaar, and Nikolaus Himmelmann, 110–181. London: Routledge.

Himmelmann, Nikolaus P. 2009. "Notes on Tagalog Nominalism." *Theoretical Linguistics* 35 (1): 115–123.

Himmelmann, Nikolaus P., and Sonja Riesberg. 2013. "Symmetrical Voice and Applicative Alternations: Evidence from Totoli." *Oceanic Linguistics* 52 (2): 396–422.

Hsieh, Henrison. 2020. Beyond Nominative: A Broader View of A'-Dependencies in Tagalog. Doctoral Dissertation, McGill University, Montreal.

Itier, César. 1997. *Parlons Quechua: La Langue du Cuzco*. Paris: L'harmattan.

Jacques, Guillaume. 2016. "Subjects, Objects and Relativization in Japhug." *Journal of Chinese Linguistics* 44 (1): 1–28. https://www.jstor.org/stable/24774959.

Jacques, Guillaume. 2021. *A Grammar of Japhug*. Berlin: Language Science Press.

Jelinek, Eloise, and Richard A. Demers. 1994. "Predicates and Pronominal Arguments in Straits Salish." *Language* 70 (4): 697–736.

Jiang, Haowen. 2016. Nominalization and Possession in Formosan Languages. Doctoral Dissertation, Rice University, Houston, Texas.

Kaufman, Daniel. 2009a. "Austronesian Nominalism and its Consequences: A Tagalog Case Study." *Theoretical Linguistics* 35 (1): 1–49.

Kaufman, Daniel. 2009b. "Austronesian Typology and the Nominalist Hypothesis." In *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, ed. by Alexander Adelaar, and Andrew Pawley, 197–226. Canberra: Pacific Linguistics.

Kaufman, Daniel. 2016. "Austronesian Clausal Constituency in a Southeast Asian Perspective." Keynote talk presented at the 26th meeting of the South East Asian Linguistics Society, Manila, Philippines, June 2016.

Kaufman, Daniel. 2017. "Lexical Category and Alignment in Austronesian." In *The Oxford Handbook of Ergativity*, ed. by Lisa Travis, Jessica Coon, and Diane Massam, 589–628. Oxford: Oxford University Press.

Kaufman, Daniel, and Victoria Chen. 2017. "Clausal Constituency and Historical Austronesian Morphosyntax." Talk presented at the 24th meeting of the Austronesian Formal Linguistic Association, University of Washington, April 2017.

Keenan, Edward L. 1985. "Relative Clauses." In *Language Typology and Syntactic Description, vol II: Complex Constructions*, ed. by Timothy Shopen, 141–170. Cambridge: Cambridge University Press.

Koptjevskaja-Tamm, Maria. 1993. *Nominalizations*. London: Routledge.

Kornfilt, Jaklin. 1997. *Turkish*. London/New York: Routledge.

Krause, Cornelia. 2001. On Reduced Relatives With Genitive Subjects. Doctoral Dissertation, MIT.

LaPolla, Randy, and Chenglong Huang. 1996. *A Grammar of Qiang with Annotated Texts and Glossary*. Ms.

Lefebvre, Claire, and Pieter Muysken. 1988. *Mixed Categories: Nominalizations in Quechua*. Dordrecht/Boston/London: Kluwer.

Lehmann, Christian. 1984. *Der Relativsatz: Typologie seiner Strukturen; Theorie seiner Funktionen; Kompendium seiner Grammatik*. Tübingen: Gunter Narr Verlag.

Li, Paul Jen-kuei. 1997. "A Syntactic Typology of Formosan Languages — Case Markers on Nouns and Pronouns." In *Chinese Languages and Linguistics IV: Typological Studies of Languages in China*, ed. by Chiu-yu Tseng, 343–378. Taipei: Academia Sinica.

Lindenfeld, Jacqueline. 1973. *Yaqui Syntax*. Berkeley/Los Angeles/London: University of California Press.

Malchukov, Andrej L. 2004. *Nominalization, Verbalization: Constraining a Typology of Transcategorial Operations*. Munich: Lincom Europa.

Manning, Christopher. 1996. *Ergativity: Argument Structure and Grammatical Relations*. Stanford, CA: CSLI Publications.

Marlett, Stephen A. 2012. "Relative Clauses in Seri." In *Relative Clauses in Languages of the Americas: A Typological Overview*, ed. by Bernard Comrie, and Zarina Estrada-Fernández, 213–242. Amsterdam/Philadelphia: John Benjamins.

Matisoff, James A. 1972. "Lahu Nominalization, Relativization, and Genetivization." In *Syntax and Semantics*, ed. by John P. Kimball, 237–257. New York: Academic Press.

Mazaudon, Martine. 1978. "La formation des propositions relatives en tibétain." *Bulletin de la Societé de Linguistique de Paris* 73: 401–414.

McGinn, Richard. 1988. "Government and Case in Tagalog." In *Studies in Austronesian Linguistics*, ed. by Richard McGinn, 457–490. Athens, Ohio: Ohio University Press.

Noonan, Michael. 1997. "Versatile Nominalizations." In *Essays on Language Function and Language Type. In Honor of T. Givón*, ed. by Joan Bybee, John Haiman, and Sandra Thompson, 373–394. Amsterdam/Philadelphia: John Benjamins.

Pan, Chia-jung. 2012. A Grammar of lha'alua, an Austronesian Language of Taiwan. Doctoral Dissertation, James Cook University.

Philippine Bible Society. 1984. *Baraan nga Biblia An Maupay nga Sumat*. Manila: Philippine Bible Society.

Post, Mark. 2011. "Nominalization and Nominalization-Based Constructions in Galo." In *Nominalization in Asian Languages*, ed. by Foong Ha Yap, Karen Grunow-Harsta, and Janick Wrona, 255–288. Amsterdam: John Benjamins.

Post, Mark, and Yankee Modi. 2011. "Language Contact and the Genetic Position of Milang (Eastern Himalaya)." *Anthropological Linguistics* 53 (3): 215–258.

Richards, Norvin. 2000. "Another Look at Tagalog Subjects." In *Formal Issues in Austronesian Linguistics*, ed. by Ileana Paul, Vivianne Phillips, and Lisa Travis, 105–116. Dordrecht: Kluwer Academic Publishers.

Riesberg, Sonja. 2014. *Symmetrical Voice and Linking in Western Austronesian Languages*. Boston/Berlin: De Gruyter Mouton.

Ross, Malcolm. 2002. "The History and Transitivity of Western Austronesian Voice and Voice Marking." In *The History and Typology of Western Austronesian Voice Systems*, ed. by Fay Wouk, and Malcolm Ross, 17–63. Canberra: Pacific Linguistics.

Ross, Malcolm. 2006. "Reconstructing the Case-Marking and Personal Pronoun Systems of Proto-Austronesian." In *Streams Converging into an Ocean: Festschrift in Honor of Professor Paul Jen-kuei Li on His 70th Birthday*, ed. by Henry Yongli Zhang, Lillian Mei-chin Huang, and Dah-an Ho, 521–63. Taipei: Institute of Linguistics, Academica Sinica.

Ross, Malcolm. 2008. "Negative Verbal Clause Constructions in Puyuma: Exploring Constructional Disharmony." In *Constructions and Language Change*, ed. by Alexander Bergs, and Gabriele Diewald, 171–194. New York: Mouton de Gruyter.

Ross, Malcolm. 2009. "Proto-Austronesian Verbal Morphology: A Reappraisal." In *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, ed. by Alexander Adelaar, and Andrew Pawley, 295–326. Canberra: Pacific Linguistics.

Sagart, Laurent. 2004. "The Higher Phylogeny of Austronesian and the Position of Tai-Kadai." *Oceanic Linguistics* 43 (2): 411–44.

Sasse, Hans-Jürgen. 2009. "Nominalism in Austronesian: A Historical Typological Perspective. Comments on Daniel Kaufman's 'Austronesian Nominalism and its Consequences'." *Theoretical Linguistics* 35 (1): 167–181.

Schachter, Paul, and Fe T. Otanes. 1982. *Tagalog Reference Grammar*. University of California Press.

Scheerer, Otto. 1924. "On the Essential Difference Between the Verbs of the European and the Philippine Languages." *Philippine Journal of Education* 7: 1–10.

Shibatani, Masayoshi. 2009. "Elements of Complex Structures, Where Recursion Isn't: The Case of Relativization." In *Syntactic Complexity: Diachrony, Acquisition, Neuro-Cognition, Evolution*, ed. by Talmy Givon, and Masayoshi Shibatani, 163–198. Amsterdam: John Benjamins.

Shibatani, Masayoshi. 2019. "What is Nominalization? Towards the Theoretical Foundations of Nominalization." In *Nominalization in Languages of the Americas*, ed. by Roberto Zariquiey, Masayoshi Shibatani, and David W. Fleck, 15–167. Amsterdam/Philadelphia: John Benjamins.

Starosta, Stanley, Andrew Pawley, and Lawrence A. Reid. 1982/2009. "The Evolution of Focus in Austronesian." In *Formosan Linguistics: Stanley Starosta's Contributions*, ed. by Elizabeth Zeitoun, 329–480. Taipei: Institute of Linguistics, Academia Sinica.

Stevens, Alan M. 1969. "Case Grammar in Philippine Languages." Paper presented at the 44th Annual Meeting of the Linguistic Society of America, San Francisco, December 1969, December 1969.

Sung, Li-May. 2011. "Clausal Nominalization in Budai Rukai." In *Nominalization in Asian Languages: Diachronic and Typological Perspectives*, ed. by Foong Ha Yap, and Janick Wrona, 523–560. Amsterdam/Philadelphia: John Benjamins.

Teng, Stacy F. 2008. *A Reference Grammar of Puyuma, an Austronesian Language of Taiwan*. Canberra: Pacific Linguistics.

Teng, Stacy F. 2009. "Case Syncretism in Puyuma." *Language and Linguistics* 10 (4): 819–844.

Teng, Stacy F., and Elizabeth Zeitoun. 2016. "Noun-Verb Distinction in Kanakanavu and Saaroa: Evidence from Pronouns." *Oceanic Linguistics* 55 (1): 134–161.

Tsuchida, Shigeru. 1976. *Reconstruction of Proto-Tsouic Phonology*. Tokyo: Tokyo University of Foreign Studies.

Tsukida, Naomi. 2009. セデック語(台湾)の文法 [Grammar of the Seediq Language]. Doctoral Dissertation, University of Tokyo, Tokyo.

Underhill, Robert. 1972. "Turkish Participles." *Linguistic Inquiry* 3 (1): 87–99. https://www.jstor.org/stable/4177687

Wang, Shan-Shan. 2004. An Ergative View of Thao Syntax. Doctoral Dissertation, University of Hawai'i at Mānoa.

Watters, David E. 1978. "Speaker-Hearer Involvement in Kham." In *Papers on Discourse*, ed. by Joseph Grimes, 1–18. Dallas: SIL and University of Texas at Arlington.

Weber, David. 1989. *A Grammar of Huallaga (Huánuco) Quechua*. Berkeley CA: University of California Press.

Wild, Ilka. 2018. Voice and Transitivity in Kanakanavu. Doctoral Dissertation, Universität Erfurt.

Wolff, John U. 1965. Cebuano Visayan Syntax. Doctoral Dissertation, Yale University.

Wolff, John U. 1973. "Verbal Inflection in Proto-Austronesian." In *Parangal kay Cecilio Lopez*, ed. by Andrew Gonzalez, 71–91. Quezon City: Linguistic Society of the Philippines.

Zeitoun, Elizabeth, and Stacy F. Teng. 2016. "The Position of Kanakanavu and Saaroa within the Formosan Languages Revisited." *Oceanic Linguistics* 55 (1): 163–198.

Zobel, Erik. 2002. "The Position of Chamorro and Palauan in the Austronesian Family Tree: Evidence From Verb Morphosyntax." In *The History and Typology of Western Austronesian Voice Systems*, ed. by Malcolm Ross, and Fay Wouk, 405–34. Canberra: Pacific Linguistics.

Zorc, David Paul. 1977. *The Bisayan Dialects of the Philippines: Subgrouping and Reconstruction*. Canberra: Pacific Linguistics.

CHAPTER 21

What's a passive in Besemah?

Bradley McDonnell
University of Hawai'i at Mānoa

In languages of western Indonesia with symmetrical voice systems, apparent passive constructions often display overlapping features with transitive P-oriented voice (PV) constructions. Drawing on a corpus of everyday conversations in Besemah, this chapter demonstrates that it is untenable to distinguish transitive PV constructions from such passive constructions, calling into question core features of both: A arguments are not obligatory in unmarked PV constructions; 'by'-phrases serve multiple functions, and it is unclear whether A demotion is in fact one of these functions; PV constructions without A arguments exhibit the same patterns of unrealized A reference found elsewhere in the language. This chapter has broader implications for the analysis of both passives and symmetrical voice alternations in the languages of western Indonesia.

Keywords: passive, voice, Malay, symmetrical voice, discourse reference

1. Introduction

Passive constructions in the world's languages have been discussed almost ad nauseam, and yet in the Austronesian languages of western Indonesia there are still fundamental disagreements about what is and what is not considered a passive construction in a number of languages, what morphosyntactic properties are fundamental to them, and what motivates their use. These debates go beyond terminology and in some cases represent profound differences in how linguists analyze the grammars of languages like Indonesian (Arka & Manning 2008; Kroeger 2014) and Acehnese (Durie 1988; Lawler 1988; Legate 2012). The complexities surrounding passive constructions often arise due to the unique nature of the voice systems in these languages, in which there are multiple apparent transitive constructions. Most often, these systems have been analyzed as so-called *symmetrical voice* systems, in which there are (at least) two transitive constructions, neither of which is considered 'basic' (Himmelmann 2005; Riesberg 2014a; Chen & McDonnell 2019; Kroeger & Riesberg 2024). In the languages of western Indonesia, symmet-

https://doi.org/10.1075/slcs.240.21mcd
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

rical voice systems typically involve a two-way distinction between A-oriented and P-oriented constructions. These constructions show little difference in terms of morphosyntactic complexity, syntactic transitivity, and frequency in discourse. Consider examples in (1)–(3) from Besemah (glottocode: pase1242), a dialect of South Barisan Malay, which is introduced in more depth in Section 2.[1]

(1) A-Voice

a. kamu asak la m-buat cuke tu.
2PL indeed PFV AV-make vinegar.sauce DEM.DIST
'you already made the vinegar sauce.'
(BJM01-158-01, 00:17:57–00:17:59, Leksi)

b. kudai aku ng-ambik=e Ci.
first 1SG AV-take=3 VOC
'I'll take it, first, Leksi.' (BJM01-158-01, 00:06:46–00:06:48, Neti)

(2) P-Voice with third person A

a. kite ni di-rekam jeme.
1PL.EXCL DEM.PROX PV-record person
'someone is recording us.' (BJM01-033-01, 00:56:45–00:56:47, Sirajudin)

b. mane cipit=e ni di-tetak=e.
which PV.new.growth=3 DEM.PROX PV-cut=3
'and he cuts its new growth.' (BJM01-011-01, 00:15:08–00:15:11, Jamisah)

(3) P-Voice with first/second person A

a. anye kabah ku=tanye-ghi.
but 2SG 1SG=PV.ask-APPL.LOC
'but I asked you.' (BJM01-008-01, 00:25:28–00:25:30, Emi)

b. 1 budak tuk n-jual Jarum tadi,
person N.LI AV-sell J. earlier
2 la kami garih-i.
PFV 1PL.EXCL PV.visit-APPL.LOC
'we already visited with the girl who sold Jarum (cigarettes) earlier.'
(BJM01-004-01, 00:49:06–00:49:11, Piter)

1. Transcription conventions follow a simplified version of Discourse Transcription (Du Bois et al. 1992; 1993). They are as follows. Each line represents an Intonation Unit (IU) with end marks representing different intonation contours: a period (.) a 'final' contour, a comma (,) a 'continuing' contour, a question mark (?) an 'appeal' contour. Dashes represent truncation of a word (en dash, –) or an IU (em dash, —). Square brackets represent overlapping speech. Short pauses are represented by (..), while timed long pauses are represented as seconds within parentheses. Colons represent lengthening. @ represents a pulse of laughter.

The examples in (1), referred to here as A-Voice (AV), privilege the A argument. The verbal predicates are prefixed with a homorganic nasal, represented by *N-*. The examples in (2) and (3), referred to as P-Voice (PV), privilege the P argument. The verbal predicate is variably marked with the prefix *di-* when A is third person, as in (2), or simply unmarked when A is first or second person (see Section 2 for a detailed description of these constructions). The privileged A and P arguments in AV and PV constructions, respectively, have been referred to by many labels: *subject* (Riesberg 2014a), *pivot* (Chen & McDonnell 2019), or even *trigger* (Cumming 1991) since it is thought to 'trigger' either AV or PV. However, in keeping consistent with previous analyses of Besemah, this privileged argument is referred to as the *primary argument* (McDonnell 2016; 2022). A defining characteristic of symmetrical voice is the transitivity of both AV and PV, evidenced by the presence of a (non-privileged/non-subject/non-pivot) core argument, which here is referred to as the *secondary argument*. In AV, this argument is a post-verbal P argument, while, in PV, secondary arguments include pre- or post-verbal A arguments.

Apparent passive voice constructions and their analyses are diverse in the languages of the region (see Kroeger & Riesberg 2024). Some languages have no passive construction whatsoever (e.g. Totoli, Riesberg 2014a), while others have a passive construction that is fully distinct from symmetrical voice (e.g. Buginese *i-* passive, Laskowske 2016). There is, however, a significant number of languages, primarily those labeled *Indonesian-type* languages (McDonnell & Chen 2022; Kroeger & Riesberg 2024), in which there is an apparent passive voice construction, but it has overlapping forms and syntactic properties with transitive PV constructions. These apparent passive constructions display properties of well-established definitions of passives, at least on the surface (Keenan & Dryer 2007; Zúñiga & Kittilä 2019; Legate 2021). That is, (i) the verbal predicate is marked, (ii) P of a transitive equivalent is the subject (or privileged syntactic argument), and (iii) A of a transitive equivalent is unrealized or occurs in a 'by'-phrase. These properties are found in the Besemah examples in (4).

(4) (Apparent) passive constructions

a. ijat=e di-jemugh.
seed=3 PV-dry
'its seeds are dried' (BJM01-093-01, 00:28:57–00:28:59, Ardin)

b. aku di-tilpun-i nga Sul kate=nye.
1SG PV-call-APPL.LOC with S. QUOT=3
'I was called by Sul, she said.' (BJM01-073-01, 01:02:38–01:02:40, Yanti)

In (4a), the same form of the PV prefix *di-* appears on the verb, but without any A argument. Thus, this example appears to be an agentless passive, at least on the surface. In (4b), the verb is again prefixed with the PV prefix *di-*, but in this case

A is preceded by a preposition *nga* 'with, by'. By all accounts, the A appears to be 'demoted' to the periphery in a 'by'-phrase (see Donohue 2007 for a similar characterization of voice in Malay/Indonesian).

However, overlapping features have led to much debate over the division between passive and transitive (PV) constructions, resulting in different analyses of the same or similar constructions in various languages of western Indonesia. In an effort to tease apart differences between transitive and passive constructions, studies have relied on a battery of syntactic diagnostics (e.g. binding, quantifier float) for proposing that one or another construction is considered a passive or transitive. Arka and Manning (2008) is a landmark article in this regard, providing a nuanced understanding of passive constructions that laid the foundation for how voice in a number of western Indonesian languages were analyzed. For example, based on evidence from binding, they find that Indonesian constructions that look like the construction in (2b) are transitive while others that look like the construction in (2a) are passive, even when the A argument is unmarked and does not occur in a 'by'-phrase.

A number of studies follow Arka and Manning's (2008) lead, describing three voices in Indonesian-type languages, although the description and categorization of each voice may differ slightly across sources (see Jeoung 2017: 20 for a fuller list of the diversity of terminology used for these voices). However, several studies have diverged from this analysis based on syntactic evidence. After reexamining Arka and Manning's binding evidence, Kroeger (2014) finds that there is no longer a syntactic asymmetry between different realizations of A in *di-* marked constructions. Based on this evidence, he concludes that all *di-* marked constructions are passive. Patrianto and Chen (2023) employ a range of syntactic diagnostics in Surabaya Javanese and argue that there is in fact no evidence for a passive construction in the language: the apparent passive prefix is in fact a third person pronominal form *di-* and all such constructions are essentially transitive constructions.

Since Shibatani's (1985) seminal article, passive constructions across the world's languages have been viewed as more or less adhering to a set of prototypical features, and even recent definitions by Zúñiga and Kittilä (2019) and Legate (2021) continue in this tradition.[2] In each of these definitions, the status of the agent plays a critical, if not the central, role. For Zúñiga and Kittilä (2019: 83), the prototypical passive involves demotion where A is 'peripheral' and 'optional', 'typically marked by a non-core case or adposition'. For Keenan and Dryer (2007), the agent is completely suppressed in the 'basic passive'. Thus, passive analyses crucially rely on 'suppression' when A is not realized, as in agentless passives, and

2. Legate (2021) uses the terms (non)canonical.

'demotion' when A is realized in a 'by'-phrase. These criteria, as it turns out, are much more complicated in the languages of western Indonesia, and specifically, in Besemah, as I show in this chapter. I discuss the complications of 'suppression' and 'demotion', each in turn, below.

The notions of suppression of A is a critical area of inquiry as it faces an often overlooked problem in the analysis of passive in most any western Indonesian language: arguments of the predicate are freely and commonly unrealized. How, then, is an agentless passive any different than a transitive construction with an unrealized A argument? Legate (2012) and Jeoung (2017) are, in fact, among the few studies to raise this issue in their respective analysis of Acehnese and Madurese passive constructions. Legate (2012), for example, proposes that if the constructions in question were indeed transitive, then the unrealized A argument is an instance of 'pro-drop'. In this case, she argues that we would expect a definite interpretation of A, but, if the construction were a passive, we would expect A to receive an existential interpretation (i.e. an indefinite 'someone/something'). To support her argument Legate (2012) provides judgements for passive examples like, 'The child was bitten, but I don't know by what.', where only the existential reading is felicitous (see Legate 2021: 163–164, for an overview of this argumentation).

Recent research in a small number of western Indonesian languages, however, has shown that what has been analyzed as 'pro-drop', 'zero anaphora', or 'elided arguments' in elicited speech is not well supported when considering language-in-use, especially conversational data (Ewing 2014, 2019, 2024, 2025; Djenar, Ewing & Manns 2018). Following work in Japanese (Nariyama 2003) and Korean (Oh 2007), Ewing and colleagues argue that the use of 'minimal structures' (i.e. predicates without arguments) is not the exception but a feature of the western Indonesian languages they have studied (see especially Ewing 2019). Rather than "speculate about putative full forms from which minimal forms are derived", Djenar et al. (2018: 112) propose that "[i]t is more productive to analyze how actual utterances are intimately linked to [...] contexts of language use." Central to this analysis is that utterances depend critically on their interactional contexts and "the constantly renegotiated intersubjective relationship that exists between interactants".

Supporting evidence for Djenar et al.'s (2018) position comes from the fact that reference in colloquial Indonesian is often indeterminate, and referents cannot always be simply 'reconstructed', as we would expect if arguments were simply dropped, elided or zero anaphora. Adopting Goffman's (1983) notion of allusion, Djenar et al. (2018: 138) show that "a heightened intersubjective attitude to the maintenance and augmentation of common ground is crucial to successful communication using minimal, allusive structures." Examples of minimal structures,

in general, and allusive reference, more specifically, also occur in Besemah, which is exemplified in the extended example in (5).

In this excerpt, a group of five young men are talking about a motorcycle race that they had attended earlier that day. After a lull in the conversation, Hendi introduces the topic of the police officer who he felt was showing off at the race earlier that day in lines 1–5, then he and Buwuh discuss the type of police officer he is in lines 6–14. While allusive reference occurs earlier on in this excerpt (e.g. line 11), line 15 is the clearest example of the sort of allusive reference Djenar et al. (2018) describe. In this line, Hendi uses the AV-marked predicate *ngaruk* 'bother' without any arguments. The referents in the free translation in English are allusive, and while it is not possible to point back to any particular previous mentions of 'the racers' or 'the spectators', none of the participants question Hendi and the conversation progresses without issue. As Ewing (2024) shows for colloquial Indonesian, participants are able to interpret these allusive referents based on their shared knowledge of the race earlier that day and the immediate context about the police at the event. This example demonstrates that successful communication proceeds despite any indeterminacy. Allusive reference in Besemah is discussed further in Section 2.3.

(5) Minimal structures and allusive reference

1. H: pelisi tadi ujuk-an=lah Gui eh.
   police earlier show.off-NMLZ=FOC G. TAG
   'those police were show-offs, right Gui.'
2. I: @@@
3. H: nik bunguk daging bange-ghan.
   N.LI fat flesh stupid-NMLZ
   'the fat one was stupid.'
4. R: @ngape?
   why
   'why?'
5. B: ui nik di-pinggir-ka=nye gale tadi.
   EXCLAM N.LI PV-edge-CAUS=3 all earlier
   'the one who tried to edge people out of the crowd.'
6. H: berim- brimob,
   TRUN riot.police
   'the riot police,'
7. I: hm.
   hmm.
   'hmm.'
8. H: pelisi tu.
   police DEM.DIST
   'the police.'

9 ame nik bepe- pikak banyak tu,
TOP N.LI TRUN pocket many DEM.DIST
'as for the one with many pockets,'

10 brimob.
riot.police
'(they) are riot police.'

11 B: begas~begas-i=nye be-
PLUR~hit-APPL=3 TRUN
'they hit (people) using'

12 tuape titu peranti die,
what DEM.DIST tool 3
'what is that, the equipment they,'

13 H: au.
yes
'yeah.'

14 B: sangi tilpun.
RECP PV.call
'call each other on.'

15 H: nanti la dide ng-(k)aruk jadi=lah.
if PFV NEG AV-bother okay=FOC
'if (the spectators) weren't bothering (the racers), okay then.'
untung jeme galak n-(t)untun.
fortunate people want AV-watch
'(it's) fortunate people wanted to watch (the race).'

16 I: au.
yes
'yeah'

(BJM01-004-01, 00:34:30–00:34:52, Hendi, Iril, Buwuh)

An analysis of passive constructions in which reference to A is allusive muddies the waters of any clearcut distinction between PV with 'pro-drop' and passive with argument suppression. It even goes so far as to call into question whether it is possible to tease apart PV constructions with allusive reference from so-called agentless passives. I return to this thorny issue in Section 3.3.

The second issue is that a core feature of passive constructions is the demotion of A into an optional, adjunct PP. In western Indonesian languages, demotion of A has been claimed even when an NP is not marked by a preposition (Arka & Manning 2008; Cole, Hermon & Yanti 2008; Kroeger 2014; Jeoung 2017). Demotion of A is analyzed as a binary, black and white parameter; A is either demoted and adjunct, or A is not demoted and remains core (Shibatani 1985; Zúñiga & Kittilä 2019; Legate 2021). However, Riesberg (2014b) questions this sort of dis-

crete analysis (see also Arka 2019, who proposes that the core-oblique distinction is gradient in nature). Through comparisons of German, Balinese, and Totoli, she proposes that the status of A in PV and passive constructions may be better regarded as gradient and ought not be analyzed as an adjunct. Ultimately, she proposes that A in 'by'-phrases are not simply adjuncts but hold an intermediate position between core and adjunct that she terms, *oblique core arguments*. This position, then, calls into question analyses that draw a sharp distinction between PV and passive, and evidence from the use of 'by'-phrases in Besemah language-in-use supports this position; there is very little functional difference between PV and apparent passive constructions with 'by'-phrases in Besemah.

Drawing on a corpus of everyday conversations, I show that the variation of overlapping patterns of transitive PV and apparent passive constructions in Besemah extends beyond what is typically described in the literature on Malay and other western Indonesian languages. Building upon my previous analysis of the two-way symmetrical voice distinction in Besemah (McDonnell 2016, 2022), I argue that drawing a sharp distinction between transitive PV constructions and passive constructions is untenable. As McDonnell (2016, 2022) demonstrates, morphosyntactic diagnostics do little to reveal any meaningful differences between secondary A arguments, and so this study looks to language-in-use to understand any such differences. While there are some examples of PV constructions that align with passives in both form and function, the vast majority of cases do not neatly map passive-like forms to passive-like functions, demonstrating instead a complex array of morphosyntactic patterns that serve diverse discourse and pragmatic purposes.

The remainder of this paper is organized as follows. Section 2 introduces Besemah and describes basic properties of symmetrical voice and allusive reference. Section 3 presents the diversity of PV constructions in Besemah, first describing PV constructions with different realizations of A (Section 3.2) followed by PV constructions without A (Section 3.3). Section 4 concludes the paper.

2. Besemah

Besemah is spoken by some 330,000 people in the highlands of southwest Sumatra, Indonesia. It belongs to a complex network of 11 named dialects, known as South Barisan Malay with an estimated 1.5 million speakers (McDowell & Anderbeck 2020). Despite the relatively large number of speakers, Besemah and South Barisan Malay more generally remain under-described. Besemah is a good example of a vernacular Malay, which is a variety that is regularly inherited from Proto Malayic; this contrast with literary Malay varieties (e.g. Standard Indone-

sian) and vehicular Malay varieties (e.g. Baba Malay). See Adelaar (2005) and McDonnell et al. (2024) for discussion of the sociolinguistic typology of Malayic languages.

2.1 Corpus

This chapter draws entirely on examples from language-in-use, more specifically a corpus of 27 everyday conversations between two to five native speakers of Besemah. In total, the corpus is just over 250,000 words and just under 95,000 Intonation Units. The documentation of Besemah, including the corpus of everyday conversations, is archived with the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC, McDonnell 2008-present).[3]

2.2 Voice and grammatical relations

Besemah shows similar morphosyntactic properties to other historically conservative varieties of Malay (McDonnell et al. 2024). Clauses consist of *predicate complex* and a primary argument: S, A in AV, and P in PV. For intransitive clauses, the predicate complex is simply the predicate. In transitive clauses, it is made up of the predicate and any secondary argument: P in AV and A in PV. As McDonnell (2016; 2022) has shown, primary arguments (S, A in AV, P in PV) behave in the same ways across clauses and different construction types. For example, they may occur either before or after the predicate complex. Secondary arguments, on the other hand, do not always behave in the same way. While A in PV is strictly adjacent to the predicate, P in AV does not have such a restriction, it is possible for adverbial elements to intervene between P in AV and the predicate, as in (7c) below. For this reason, secondary arguments are often referred to in this chapter as secondary A or secondary P arguments.

Aside from a handful of unmarked transitive verbs (e.g. *ade* 'have', *ghulih* 'get'), transitive predicates are obligatorily prefixed with the A-Voice (AV) *N-*, but are optionally prefixed with *di-* in P-Voice (PV) when the secondary A argument is third person or is not realized. Otherwise, the verb is unmarked in PV constructions. The word order properties above are exemplified in intransitive (6), AV (7), and PV when prefixed with *di-* with a third person A (8) and when un-

3. The sessions that are included in this corpus include the following: BJM01-001, BJM01-002, BJM01-004, BJM01-008, BJM01-010, BJM01-011, BJM01-015, BJM01-028, BJM01-029, BJM01-032, BJM01-033, BJM01-035, BJM01-041, BJM01-043, BJM01-071, BJM01-073, BJM01-081, BJM01-086, BJM01-088, BJM01-093, BJM01-098, BJM01-116, BJM01-125, BJM01-129, BJM01-130, BJM01-131, BJM01-158.

marked with a first/second person A (9). In these examples, the pre-predicate complex primary arguments are in (a) while the post-predicate complex primary arguments are in (b).

(6) Intransitive variable word order
   a. aku duduk di depan.
      1SG sit LOC front
      'I sat in the front.' (BJM01-008-01, 00:19:35–00:19:37, Emi)
   b. dindak duduk adik tu.
      not.want sit younger.sibling DEM.DIST
      'the younger sibling didn't want to sit.'
      (BJM01-131-01, 01:08:57–01:08:59, Mala)

(7) AV variable word order
   a. aku la udim .. ng-ambik mulan kawe di situ eh?
      1SG PFV finish AV-take seedling coffee LOC there TAG
      'I already took the coffee seedlings there, right?'
      (BJM001-001-01, 00:20:50–00:20:52, Yawan)
   b. m-beli rukuk die.
      AV-buy cigarette 3
      'he bought the cigarettes.' (BJM01-008-01, 00:21:14–00:21:16, Emi)
   c. aku laju m-beli agi bakso nga bik Nis kate=nye.
      1SG then AV-buy again meatball with aunt N. QUOT=3
      'then I bought the meatball with Nis, she said.'
      (BJM01-043-01, 00:45:51–00:45:54, Leksi)

(8) PV with *di-* variable word order
   a. kite ni di-rekam jeme.
      1PL.INCL DEM.PROX PV-record people
      'someone recorded us.' (BJM01-033-01, 00:56:45–00:56:47, Sirajudin)
   b. kunake di-gaghi jeme endie die tu:.
      but PV-visit people indeed 3 DEM.DIST
      'but people really visited here.'
      (BJM01-071-01, 00:39:38–00:39:41, Musrani)

(9) PV with first/second person A variable word order
   a. anye kabah ku=tanye-ghi.
      but 2SG 1SG=PV.ask-APPL.LOC
      'but I asked you.' (BJM01-008-01, 00:25:28–00:25:30, Emi)
   b. la nak kami susul Iwin.
      PFV want 1PL.INCL PV.follow I.
      'we already want to follow Iwin.' (BJM01-130-01, 00:32:02–00:32:05, Pebri)

PV constructions show additional complexities and vary along two dimensions: (i) the presence or absence of the prefix *di-* 'PV' and (ii) the realization of A, as a verb adjacent pronominal clitic, verb adjacent noun, 'by'-phrase headed by *nga* 'with' or *li* 'by', a combination of pronominal clitic and 'by'-phrase. The examples thus far have shown patterns that are commonly described in Malay and other western Indonesian languages.[4] However, the patterns below are either less commonly found in Malay and other western Indonesian languages or have simply not been described in many of these languages. Verbs that are unmarked or prefixed with *di-* can occur with all possible realizations of A arguments, including unrealized A arguments. The only restriction is that *di-* cannot combine with a first or second person pronominal clitic. The examples below demonstrate the variability of *di-* with pairs of constructions in which A is a third person enclitic (10) or a verb adjacent noun (11). The examples in (a) occur with the *di-* prefix, and in the examples in (b) the verb is unmarked.

(10) PV third person A enclitic, variable *di-*

a. kami dide: di-ajar-i=nye.
1PL.EXCL NEG PV-chase-APPL.LOC=3
'he didn't chase us.' (BJM01-029-01, 00:50:13–00:50:15, Neti)

b. racun tumpah-ka=nye.
poison PV.pour-CAUS=3
'he poured out the weed killer.'
(BJM01-093-01, 00:47:59–00:48:00, Gunawan)

(11) PV with nominal A, variable *di-*

a. 1 deghian bantal,
durian pillow
2 besak panjang tu,
big long DEM.DIST
3 la di-pajuh tupai.
PFV PV-eat squirrel
'a squirrel ate the big, long durian.'
(BJM01-131-01, 00:29:09–00:29:11, Mala)

b. dide ka gigit-i agas aku
NEG FUT PV.bite-APPL mosquito 1SG
'mosquitos won't bite me.' (BJM01-029-01, 00:01:46–00:01:48, Neti)

4. There are no apparent restrictions on the use of either *nga* 'with' or *li* 'by' in PV constructions. These prepositions have different meanings when used in other constructions and combine differently with clitic pronouns (i.e. *li* 'by' may combine with *=ku* '1SG' and *=(ny)e* '3', while *nga* 'with' may not).

The same is true in apparent passive constructions when A occurs in a 'by'-phrase (12) or when A is not realized (13). Again, the examples with the *di-* prefix are in (a) while the examples with the unmarked verb are in (b). The examples in (b) are especially surprising. For (12b), apparent unmarked passives are not attested in Malay or other western Indonesian languages. For (13b), there are two potential analyses, which is the reason for the two different free translations for this example. The first analysis is that it is an unmarked passive that can even occur without a 'by'-phrase. The second analysis is that it is a PV construction that does not obligatorily realize an A argument, which is surprising because unmarked PV (i.e. Object Voice) constructions are assumed to obligatorily realize A arguments (Kroeger 2014). I return to this point in Section 3.

(12) PV with 'by'-phrase, variable *di-*

a. di-ambik li mamak Aji.
PV-take by uncle A.
'(it) was taken by uncle Aji.' (BJM01-001-01, 00:48:26–00:48:29, Juria)

b. ambik=lah li endung=e mubil.
PV.take=FOC by mother=3 car
'his mother (should) take the car.'
(BJM01-011-01, 00:29:37–00:29:41, Jamisah)

(13) PV without A, variable *di-*

a. sate udim di-putigh,
after finish PV-pick
'after (the vanilla) was picked,'
(BJM01-010-01, 00:35:32–00:35:34, Burhimin)

b. dide nak jagal
NEG want PV.chase
'(they) shouldn't be chased'
'(we) shouldn't chase (them)' (BJM01-041-01, 00:28:26–00:28:28, Iril)

Finally, there are PV constructions that express A coreferentially, with a combination of pronominal clitic and a 'by'-phrase, most commonly a third person enclitic *=nye* and a 'by'-phrase with an NP, as in (14). These constructions again may occur with or without the prefix *di-*.

(14) PV with coreferential *=nye* and 'by'-phrase

a. ade dide die di-pantau=(ny)e agi li endung Erda?
exist NEG 3 PV-call=3 again by mother E.
'was she (or was she) not called by Erda's mom?'
(BJM01-008-01, 00:19:53–00:19:56, Ina)

b. 1 tanye-ka=nye agi,
ask-APPL=3 again,
'he asked again,'
2 nga bapang Gitah rasan=(ny)e.
with father G. agreement=3
'Gitah's father, (about) the agreement.'
(BJM01-011-01, 00:02:03–00:02:05, Rili)

It is also possible for these PV constructions to involve two coreferential pronouns, as in (15). In the corpus, there are only instances of first person singular (15a) and third person singular or plural (15b–d) coreferential arguments. The examples below also demonstrate how the verb unsurprisingly is able to be variably marked with *di-* when A is a third person referent. In Section 3.2, I discuss these constructions in more detail, including some of their functional properties.

(15) PV with coreferential pronouns in 1st and 3rd person with variable *di-*
a. ku=katup liku duaghe.
1SG=PV.close by.1SG door
'I myself close the door.' (BJM01-015-01, 00:25:55–00:25:58, Rumsiah)
b. 1 adi anak=(ny)e masih kecik~kecik tadi,
so child=3 still DISTR~small earlier
'so their kids (who) were still small then,'
2 batak-i=nye gale li die.
PV.bring-APPL.LOC=3 all by 3
'were all brought by them.'
(BJM01-011-01, 00:58:14–00:58:16, Rili)
c. di-pupuk=(ny)e gale: linye.
PV-fertilize=3 all by.3
'he fertilized all (of it).' (BJM01-033-01, 00:06:50–00:06:52, Muliati)

2.3 Allusive reference and the challenge of unrealized arguments

As mentioned in Section 1, in Besemah, as in many of the languages of western Indonesia, arguments of the predicate are commonly unrealized. Rather than viewing the unrealized arguments as being 'dropped', 'elided' or otherwise 'omitted', I follow a series of studies primarily investigating everyday conversations in colloquial Indonesian that view such minimal structures as complete utterances that rely on intersubjectivity, recognizing that 'all utterances are dependent on the contexts in which they are produced, including their discourse context, the wider social-cultural context of a speech event and crucially the constantly renegotiated intersubjective relationship that exists between interactants' (Djenar et al.

2018: 112). This process is 'inherently collaborative, requiring as it does active work on the part of both speaker and hearer to arrive at an interpretation that is shared' (Djenar et al. 2018: 107; citing Fox 1994: 2). In this section, I provide several examples of minimal structures from Besemah conversation demonstrating how allusive reference often results in fluid and indeterminate referents. This section lays the foundation for further discussion of PV constructions without A, apparent agentless passives, in Besemah described in detail in Section 3.3.

In the extended excerpt in (16) below, Sira, his wife, Mariati, and his older sister, Muliati, are talking while drying rice. In the first two lines, Sira asks his sister about *padi* 'husked rice' that she's storing under her house. It becomes clear later that Sira intends to ask her for this husked rice, so he can make seedlings from it, as an earlier attempt to grow rice with different husked rice had failed. Mariati begins in line 4 to explain what happend to their husked rice in this failed attempt.

(16) Minimal structures and allusive reference

1 S: tuk ng-ambik nik kamu madake,
N.LI AV-take N.LI 2PL formerly
'the one (you) took, yours, back then,'

2 masih di bawah ghumah eh?
still LOC below house TAG
'(it's) still below the house, right?'

3 Ma: kalu: e-
probably TRUN
'probably'

4 la nak uni-ka=nye [nga bapang Ap][$_2$ip.]
PFV want PV.seedling-CAUS=3 with father A.
'Apip's dad, he already wanted to make (husked rice) into seedlings.'

5 S: [dindak agi]
not.want again
'(they) didn't anymore'

6 [$_2$din][$_3$dak agi tumb]uh.
not.want again grow
'(they) didn't grow anymore.'

7 [$_3$dindak tumbuh.]
not.want grow
'(they) didn't grow.'

8 Mu: m:
hmm
'hmm.'

9 Ma: padi buku itam madake.
husked.rice joint black formerly
'black joint rice back then.'
10 S: ni- bu- [ah- pagi anilah-]
TRUN TRUN TRUN tomorrow TRUN
11 Ma: [ayik kegh][$_2$uh],
water clean
'clean water,'
12 S: [$_2$ay]ik keghuh ni=lah.
water clean DEM.PROX=FOC
'the clean water.'
13 Mu: e- e'e.
TRUN uhuh
'uhuh.'
14 Ma: ka n-(t)anam padi itu eh?
FUT AV-plant husked.rice DEM.DIST TAG
'(we were) going to plant the rice, right?'
(BJM01-033-01, 00:41:28–00:41:41, Sirajudin, Mariati, Muliati)

This short dialogue demonstrates some of the minimal structures that Djenar et al. (2018) describe, with some more obvious instances of allusive reference to 'husked rice' in lines 4–7, which has been provided in parentheses in the free translation. In line 4, Mariati begins tracking 'husked rice' as the allusive referent through the use of the PV transitive verb *unika* 'to make (something) into seedlings', which is the combination of the causative/applicative suffix *-ka* and the nominal root *uni* 'seedling'. In lines 5–7, Sirah and Mariati employ negated intransitive predicates without any explicit argument, allowing Muliati to track the same allusive reference. While the allusive reference for *padi* 'husked rice' is made explicit in line 9 with additional information on the particular type (or even brand) of husked rice it is, and again referenced in line 14, the A argument of AV transitive verb *nanam* 'plant' in line 14 is allusive. What is particularly relevant here is that there is no previous mention of A, yet it is clear from context, at least clear enough for the conversation to progress (see Ewing 2024).[5] How-

5. Hendi Feriza, a native speaker of Besemah and the person who transcribed and translated this recording, included in parenthesis the first person plural exclusive pronoun *kami* 'we' in the Indonesian translation. This inference is based on knowledge of Sira and Mariati in particular and farming practices in general where Sira, Mariati and potentially other family or friends planned to help plant the rice in the fields.

ever, as Djenar et al. (2018) point out, it is not strictly possible to 'reconstruct' the referent. Would it in fact be the pronoun *kami* '1PL.EXCL' (as the free translation suggests) or Mariati's husband *bapang Apip* 'Apip's father' (as mentioned in line 4)? In which position would we reconstruct the referent, preverbally or after the predicate complex (see Section 2)? The larger point here is that when Mariati produces an utterance as in line 14 without a primary A argument, Sira (her husband) and Muliati (her sister-in-law) are able to infer an allusive referent that is enough for the interaction to progress, despite any referential indeterminacy. The participants' shared knowledge of rice-planting in general and each other more specifically allows them to maintain intersubjective alignment through the use of minimal structures.

3. Passives in Besemah?

In this section, I describe the wide range of PV constructions in Besemah along two dimensions: (i) verbal marking and (ii) realization of A. As I have shown in the previous section, verbs are prefixed with *di-* or unmarked, and A is unexpressed or expressed as a verb adjacent pronoun, a noun, a 'by'-phrase, or a combination of verb adjacent pronoun and 'by'-phrase. As each type of verbal marking may combine with each type of realization of A, there are in principle ten possible PV constructions, and in fact all ten are attested in the corpus. At the same time, in investigating a subset of the corpus, it is clear that they are not equally frequent. Some constructions are quite rare. In this section, I exemplify each of these constructions both in terms of form as well as any salient functional properties. Several of the less frequent constructions appear to be restricted in the contexts in which they are employed. Based on this discussion, I argue that it is untenable to neatly separate PV and apparent passive constructions in Besemah. In so doing, I also reveal unexpected properties of both verbal marking and the realization of A in PV. More specifically, I show the following:

1. Unmarked PV constructions in Besemah do not obligatorily require secondary A arguments, contrary to what has been claimed for Object Voice constructions in Malay and other western Indonesian languages (Arka & Manning 2008; Kroeger 2014; Kroeger & Riesberg 2024). This analysis, however, raises a second issue that the unmarked PV may also be analyzed as a passive without any verbal marking, or what Zúñiga & Kittilä (2019) refer to as an *uncoded passive*.
2. The use of 'by'-phrases in PV constructions is uncommon in discourse, and the degree to which 'by'-phrases serve any sort of demoting function is

unclear. In some cases, A in a 'by'-phrase displays signs of demotion (e.g. marking ephemeral, non-topical NPs), but in other cases it does not (marking persistent, topical NPs). 'By'-phrases commonly serve as a syntactic position for 'definite' and 'heavy' NPs that cannot appear in the restricted verb-adjacent positions in PV constructions, which is generally restricted to nouns without modifiers, determiners, or other elements. This syntactic restriction means that 'by'-phrases serve as the only available position for A NPs in PV constructions.

3. On the surface, PV constructions without an A argument, particularly those prefixed with *di-*, appear to be similar to agentless passive constructions. However, these constructions display considerable variability that complicates any sort of straightforward passive interpretation. First, while *di-* appears to be frequently correlated with PV constructions without A, unmarked verbs in such PV constructions are also common and show no discernable differences. Second, the unrealized A, as an instance of allusive reference, varies significantly in how it is interpreted by participants, from clearly defined specific referents to general, indefinite referents and everything else in between. That is, allusive reference of A in PV constructions appears to behave in the same ways it does in any other construction, calling into question the special status of A 'supression' under a passive analysis (see e.g. Legate 2012).

3.1 Overall patterns of PV constructions

To gain a broader understanding of PV constructions, I present a breakdown of the 400 instances of PV constructions in four conversations from McDonnell (2016, Chapter 7), which is a subset of the larger corpus.[6] Of the 400 PV constructions, A is unrealized in over half of all cases (211 total instances; 53%) or realized as a pronoun in over a third of all cases (145 total instances; 36%). The other realizations of A, as a noun, a 'by'-phrase, or a combination of pronoun and 'by'-phrase, are not as common in these conversations; taken together they make up just over 10% of the total instances. PV constructions with and without prefixes on the verb are roughly equal; verbs prefixed with *di-* make up a little less than half of all instances (180 in total; 45%), unmarked verbs making up the remaining 220 total instances (55%). Figure 1 presents a detailed breakdown of the frequency of the different PV constructions according to (i) the different realizations of A described in the previous section, (ii) the person and number of A (i.e. whether it is first/second person or third person), and (iii) the presence of the *di-* prefix.

6. The four conversations are BJM01-004-01, BJM01-008-01, BJM01-010-01, BJM01-011-01. See McDonnell (2016: 205–209) for a description of these conversations and the distribution of PV constructions.

Beginning with the least frequent realizations of A, nouns, 'by'-phrases, and coreferential pronouns occur with both *di-* prefixed and unmarked verbs. In unmarked PV constructions involving 'by'-phrases, first and second person A is comparable to third person A. However, these constructions are infrequent, so it is difficult to draw any conclusions. In PV constructions with unmarked verbs that occur with a pronoun, the majority are first or second person (78 instances). Recall, first and second person pronouns do not occur with verbs prefixed with *di-*. Comparing only the third person unmarked and prefixed verbs with third person A pronouns, there is a roughly equal distribution (35 ~ 32 instances). There is a sizable discrepancy between unmarked and *di-* prefixed verbs when A is unrealized; it is more common for the prefix *di-* to occur when A is unrealized. While this may lend some support to a passive analysis of *di-*, there are plenty of instances where A is unrealized and the verb is unmarked. It is also noteworthy that when A is realized, *di-* is equally likely to occur regardless of whether A is a pronoun (32 instances with *di-* prefixed verb, 35 instances with unmarked verb) or a noun (8 in each type of construction).

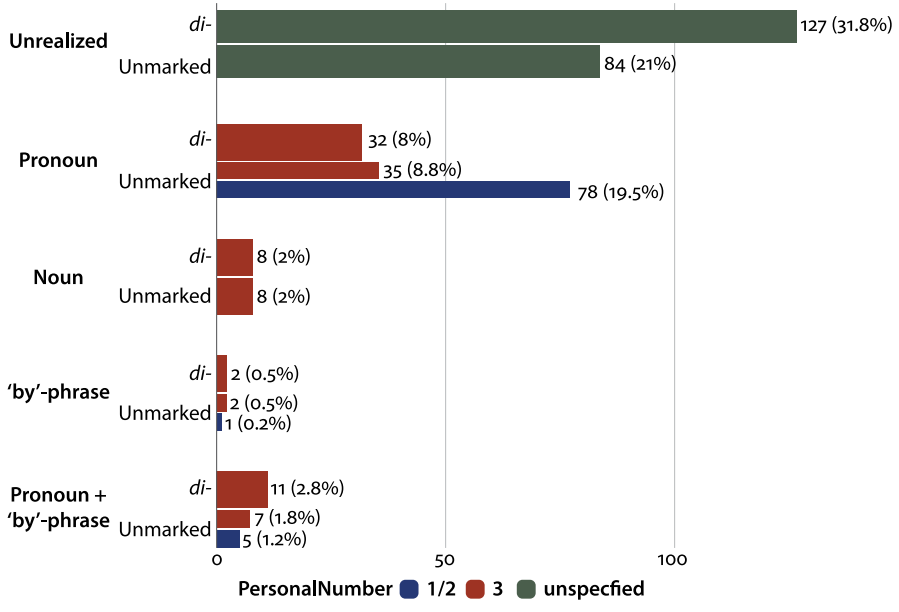


Figure 1. The frequency of the different realizations of A in PV constructions

The realizations of A in PV reflect how A is realized more generally, that is, in both AV and PV. Compare the frequencies of the realizations of A in Figure 2 from the same four Besemah conversations from McDonnell (2016, Chapter 7). The realizations of A in AV are shown in (a), while the realizations of A in PV

are shown in (b). The PV constructions are the same 400 instances in Figure 1, while the AV constructions include 499 instances. What is striking about the plots below is that the proportion of A when it is unrealized or a pronoun is comparable (51.7% in AV ~ 52.8% in PV and 35.9% in AV ~ 36.2% in PV, respectively). If we compare the remainder of forms (i.e. nouns and forms involving 'by'-phrases in PV, NPs or forms involving NPs in AV), the proportion is again comparable (12.4% in AV ~ 11% in PV). These patterns reveal that A in both AV and PV show roughly the same patterns of realization. There is nothing particularly marked about PV constructions.

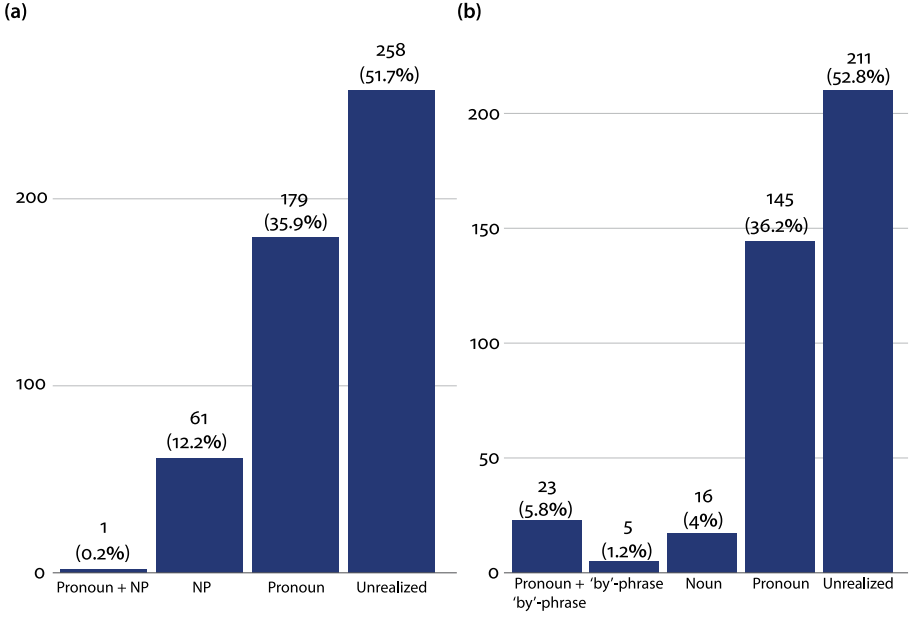


Figure 2. The frequency of the different realizations of A in AV (a) and A in PV (b)

In the remainder of this section, I explore each construction in turn drawing on the larger corpus of Besemah conversations beginning with PV constructions with A (Section 3.2), followed by constructions without A (Section 3.3). Because the focus is on constructions that have overlapping features with passives, I do not discuss constructions in which A is a pronoun in detail. As I have shown in previous work, these constructions are transitive with all of the hallmarks of transitive constructions (McDonnell 2016). Each subsection is organized by the realization of A and both unmarked and marked verbal constructions are described within each subsection.

3.2 PV constructions with A

PV constructions with a pronominal A argument are the transitive constructions typically described in Malay and other western Indonesian languages. The majority of the examples presented in the sections above all fall within this category. In the subset of Besemah conversations in Figure 1, it is common for unmarked PV constructions to occur in first, second and third person. When A is a third person referent, the verb may be unmarked or prefixed wtih *di-*, as shown in Section 1. What is relevant to the present discussion is that it is equally likely for *di-* to be present (or absent) when A is expressed as a third person pronoun or a verb adjacent noun. It is not uncommon to see such patterns in discourse where both the unmarked verb and the prefixed verb is used within the same sequence. Consider the example in (17).

(17) Unmarked PV and PV with prefix *di-*

1 R: di-tumis-keghing-ka=nye bum[bu lengkap],
PV-stir.fry-dry-CAUS=3 spice complete
'she stir fried (them) dry, full of spices,'

2 D: [kiung tu]?
snail DEM.DIST
'the snails?'

3 R: au tumis-keghing-ka=[nye.]
yes PV.stir.fry-dry-CAUS=3
'yeah, she stir fried (them) dry.'

4 D: [au]::.
yes
'yeah.'

(BJM01-011-01, 00:09:45–00:09:50, Rili, Dewi)

Eades (2005) describes something similar for the prefix *i-* in Gayo, a Sumatran language of North Sumatra, but he describes it as '*i-* dropping'. In Besemah, such a characterization is not fitting as verbs commonly occur without *di-* even upon first mention in the discourse. While uncommon, nouns similarly may appear adjacent to the verb, which may be unmarked or prefixed with *di-*, as in (18). By far, the most common noun is *jeme* 'people' (18a–b) or 'someone' (18c–d).

(18) PV verb adjacent noun *jeme* 'people'

a. mangke aku di-berenti-ka jeme.
thus 1SG PV-stop-CAUS people
'so people stopped me (from being in the group).'
(BJM01-008-01, 00:09:04–00:09:06, Emi)

b. empai nak bebak jeme.
REC.PST want PV.tear.down people
'people just now want to tear (the house) down.'
(BJM01-028-01, 01:58:03–01:58:05, Wulan)

c. di-pinjam jeme kate=nye.
PV-borrow people QUOT=3
'he said, someone borrowed (it).'
(BJM01-004-01, 00:09:12–00:09:13, Hendi)

d. aku la ghabal-i jeme dumpit.
1SG PFV PV.pickpocket-APPL.LOC people wallet
'someone already pickpocketed me'
(BJM01-004-01, 00:59:56–00:59:57, Piter)

There are, however, other nouns that may occur in these constructions, as in (19).

(19) PV with other verb adjacent nouns

a. 1 iku- guk ini=nye la,
TRUN side FILL=3 PFV
2 .. pajuh kucing
PV.eat cat
'a cat ate this side of the whatchamacallit (fish)'
(BJM01-035-01, 00:38:26–00:38:29, Ishan)

b. la luk jagal antu kami
PFV like PV.chase ghost 1PL.EXCL
'it was like we were chased by a ghost.'
(BJM01-029-01, 01:08:38–01:08:40, Liah)

c. 1 bini Budi,
wife B.
2 dik di-ajung Budi.
NEG PV-order B.
'Budi's wife wasn't allowed by Budi (to go).'
(BJM01-008-01, 00:07:21–00:07:23, Kudar)

Beginning with Arka & Manning (2008), the analogous construction in Indonesian, which requires the prefix *di-*, is characterized as 'dropping' the preposition 'by', thus it is analyzed as the same as the PV with a 'by'-phrase. This analysis has been extended to other western Indonesian languages (see e.g. Jeoung 2017). In

Besemah, however, the two positions are not equivalent. First, as mentioned in Section 2.2, McDonnell (2022) demonstrates that nouns that are adjacent to verbs are not able to be quantified by adverbial universal quantifiers in the same way that NPs in 'by'-phrases can be quantified in PV constructions. Second, in the verb adjacent position in PV constructions we see, with few exceptions, a single noun without any modifiers or even a determiner; these constructions seem to even disprefer proper nouns, although there are a few counterexamples. Rather, nouns with modifiers, determiners, or other elements occur in 'by'-phrases (see discussion below). That is, the post-verbal position in PV is restricted to nouns, and NPs appear in 'by'-phrases, whether the verb is unmarked or prefixed with *di-*. One would need to posit that the preposition is 'dropped' or is optional only in the case that A is a noun and not if it is a NP. However, I view these as two different PV constructions, a 'by'-phrase construction and a verb adjacent noun construction, for the following reasons. First, a systematic optional preposition is not found in any other constructions in Besemah, to my knowledge, so there is no reason to posit one here. Second, just as there is no reason to assume that speakers are 'omitting' arguments, there is also no reason to believe they are 'omitting' prepositions either.

Despite syntactic differences, the constructions in (18) and (19) above share particular discourse features with prototypical passive constructions, as outlined in studies by Givón (1983, 1994), wherein A (typically in a 'by'-phrase) lacks pragmatic prominence. Givón (1983) operationalizes pragmatic prominence using *referential distance* (i.e. the distance typically counted in clauses between a referent and its previous mention) and *topic persistence* and (i.e. the degree to which a referent persists in the discourse counted in subsequent mentions). Kazenin (2001: 67) summarizes it nicely, "[i]t turns out that in the passive voice, the actor is systematically characterized by high referential distance and low topic persistence…". The example in (20) where the noun A *bidan* 'midwife' is in a 'by' phrase is similar to the examples in (18) and (19) above; the noun is not previously mentioned (i.e. high referential distance) or mentioned again (i.e. low topic persistence).

(20) PV with 'by'-phrase
1 adak di-perikse-ka nga bidan,
when PV-check-APPL with midwife
'when she was checked by the midwife,'
2 dide nge-ngape,
NEG why
'there wasn't anything wrong,'
3 asak la POSYANDU,
when PFV birth.center
'when (she) went to the birthing center,'

4 la di-suntik~suntik,
PFV PV-PLUR~shot
'she was given shots,'

(BJM01-011-01, 00:34:17–00:34:22, Rili)

PV constructions with a verb adjacent noun actually resemble what Zúñiga and Kittilä (2019: 224–225) refer to as *indefinite subject constructions* (or what Haspelmath 1990 refers to as *generalized subject constructions*), which they show can develop into 'uncoded passives' or even bona fide passives (see Zúñiga & Kittilä 2019: 188–189, for discussions of 'uncoded passives').

When A is an NP or a proper noun, in almost every case, it occurs in a 'by'-phrase, as in (21). In a very small number of cases, a proper noun consisting of a single name may appear adjacent to the verb, but it is more common to occur in a 'by'-phrase. If it is more than a single word, then it almost always occurs in a 'by'-phrase.

(21) PV with A as an NP in 'by'-phrase

a. dang s(e)-aghi-an itu di-adang nga jeme Lintang.
PROG one-day-NMLZ DEM.DIST PV-block with people L.
'on that day, people from Lintang blocked (them).'

(BJM01-029-01, 00:49:21–00:49:24, Liah)

b. dide: .. (H) di-jual-ka nga bapang Apip ni.
NEG PV-sell-APPL with father A. DEM.PROX
'Apip's father didn't sell (them).'

(BJM01-033-01, 00:59:07–00:59:11, Mariati)

However, these 'by'-phrases do not necessarily fall in line with prototypical discourse features of passive constructions, summarized in the quote by Kazenin (2001) above. Compare the example of the more prototypical passive in (20) above with the examples in (22) below.

(22) PV with A as an NP in 'by'-phrase

1 K: n-(t)urai[=(ny)e nagh.]
av-imitate=3 really
'(she) is really imitating him.'

2 W: [enduk ter]ang=(ny)e.
exclam clear=3
'wow, how clear.'

3 (0.3)

4 R: a[h?]
hah
'hah?'

5 K: *[De]la:*
D.

6 (1.5)

7 W: m ninik be-uban-an.
hmm grandma mid-gray.haired-nmlz
'hmm, grandmother has gray hair.'
(2.8)

8 K: di-tiru nga Dela ni.
pv-imitate with D. dem.prox
'Dela is mimicking (him).'

(BJM01-028-01, 01:21:05–01:21:15, Kus, Wulan, Royani)

In this example, Wulan gets up to look through the viewfinder of the video camera. When he does, Dela mimics the way that he is sitting in the video. Kus, comments on this in line 1 using an AV-marked verb with a post predicate complex reference to Dela in line 5. Kus again in line 8 uses a synonymous verb but this time in PV with the NP *Dela ni* in a 'by'-phrase. Dela, as a participant in the conversation and having just been mentioned, is clearly not similar to A in prototypical passives discussed above.

Thus far, when A is in a 'by'-phrase, the verb has been prefixed with *di-*, and indeed when A is an NP (and thus also third person), the vast majority of these constructions are prefixed with *di-*. There are, however, a few cases where the verb is unmarked and A is a noun or NP in a 'by'-phrase, as in (23) and (24). These are exceptional. In the majority of cases, the verb is prefixed with *di-*, although the use of a 'by'-phrase is not altogether common, as Figure 1 shows.[7]

7. While none of the examples above are imperative constructions, the use of the enclitic *=lah* foc in (24) is noteworthy as it often occurs on verbs in imperative constructions.

(23) Unmarked PV with 'by'-phrase
empat ribu jual-ka li jeme
four thousand PV.sell-APPL by person
'the person sold (it) for four thousand.'
(BJM01-010-01, 00:14:09–00:14:11, Aripin)

(24) Unmarked PV with 'by'-phrase
1 ah kilik baih.
EXCL PV.cradle just
'ah, just pick (her) up'
2 kilik=lah nga endung=(ny)e nah.
PV.cradle=FOC with mother=3 now
'his mother will pick her up, see.'
(BJM01-043-01, 01:00:32–01:00:35, Neti)

As we saw at the beginning of this section, the same alternation between unmarked and prefixed PV verbs occurs within the same sequence. Consider the example in (25), which occurs shortly after (24). While the verbal root and the expression of the 'by'-phrase stays the same, the verb in the PV construction is prefixed with *di-*:

(25) Alternation with unmarked and prefixed PV with 'by'-phrase
1 ui dide nak badah bubuk ige.
EXCLAM NEG want place sleep exceed
'hey, (you) don't need to (mess up) the bed.'
2 ka di-kilik nga endung dengah.
FUT PV-cradle with mother 2SG
'your mother will hold (her).'
(BJM01-043-01, 01:01:17–01:01:21, Leksi)

In the vast majority of constructions with unmarked verbs and a 'by'-phrase, A is a first or second person pronoun, as in (26).

(26) PV imperative with second person pronoun in 'by'-phrase
1 bayang-ka li kabah Lik eh?
PV.imagine-CAUS by 2SG R.VOC TAG
'imagine (it), Rafles, right?'
2 sandi sini ke jungut dinding tu=lah.
from here to end wall DEM.DIST=FOC
'from here to the end of the wall.'
3 empai udim setar.
REC.PST finish start
'right after the starting line.'

4 m-(p)ulik.
AV-flip
'(he) flipped'

(BJM01-004-01, 00:44:42–00:44:49, Piter)

When A is a second person pronoun, the PV construction is often an imperative. In (26), immediately after the imperative in line 1, Piter launches into a story about a motorcycle race. The second person pronoun serves to directly address and involve the addressee in the unfolding action, either as the one who is requested to carry out the action or the one whose attention is requested, as in (26).

In a small number of examples, A is a first person pronoun, but these PV constructions are not imperatives and do not appear to carry any other sort of pragmatic or semantic effects, as in (27). In both cases, there is little discernable difference in the use of these PV constructions and those with verb adjacent pronouns above.

(27) Unmarked PV with first person pronoun in 'by'-phrase
1 ke-ghase-ghan nagh,
NMLZ-feel-NMLZ really
'(I) really felt (it),'
2 gha=itu nak me-ghabal dumpit,
person=DEM.DIST want AV-feel wallet
'that person wanted to steal (my) wallet,'
3 langsung celpik nga aku,
directly block with 1SG
'I right away blocked (it),'
4 ku=tulih,
1SG=PV.turn.head
'I turned my head,'
5 m-(p)inggir-ka jiku.
AV-edge-CAUS QUOT.1SG
'move to side, I said.'

(BJM01-004-01, 01:00:08–01:00:12, Piter)

Thus, in these kinds of examples, there is a fairly strong correlation between the absence of the prefix *di-* and the person of A in PV constructions: when the verb is unmarked, A is most often a second person pronoun but also occurs as a first person pronoun. However, recall from Section 2.2 that in PV constructions where A is expressed with a combination of a pronoun and 'by'-phrase there is a gap in second person and first person plural forms in the corpus. One explanation for this gap is that these forms are simply expressed with a 'by'-phrase whereas first

person singular and third person forms may be expressed with a combination of pronoun and 'by'-phrase.

McDonnell (2016) and Chen and McDonnell (2019) discuss such examples of coreferential A arguments in PV constructions wherein A may be realized as a pronoun and expressed within a 'by'-phrase. These realizations of A are much more common than A being expressed in a 'by'-phrase alone. When A is third person, these constructions may be best described as an instance of anti-topicalization, wherein the speaker is resolving any referential ambiguity or re-estalishing/confirming a referent (Chafe 1976). The example in (28) demonstrates antitopicalization. In line 1, Jamisah asks Rili about an arrangement that she and her husband had previously agreed to wherein they would plant crops in someone else's field. Once they harvest the crops, Rili and her husband would split the proceeds with the owner of the land. Up until this point Rili's husband has been a topic of conversation, but is not singled out as a referent until line 13 as the third person *=nye* A encliticized to the PV transitive verb *tanyeka* 'ask about' and again in the next line with a coreferential 'by'-phrase *nga bapang Gitah* 'by Gitah's father'. The coreferential 'by'-phrase appears to serve two purposes, (i) to re-establish Rili's husband as a referent and (ii) to resolve potential ambiguity with *=nye* from line 6 'they didn't return (it)', referring to the people who owned the land. As Figure 1 demonstrates, the combined use of the pronoun and the 'by'-phrase is more common than the 'by'-phrase alone.

(28) Coreferential A in PV constructions with *=nye* + 'by'-phrase

1 J: kamu n-(t)ebas madake?
2PL AV-plow previously
'you all plowed (the land) back then?'
(0.6)

2 R: laik-
TRUN

3 .. urung kami t[u].
cancel 1PL.EXCL DEM.DIST
'we canceled (it).'

4 J: [u]rung,
cancel
'(you all) canceled,'
(0.3)

5 anu=nye luk mane mudal=(ny)e?
FILL=3 like which capital=3
'the whatchamacallit how about the capital?'
(0.5)

6 R: dide di-balik-ka=nye
NEG PV-return-CAUS=3
'they didn't return (it),'
(0.5)
7 basekah titu madake,
TOP DEM.DIST formerly
'as for that (situation) back then,'
8 antakkah n-(t)anam e[h]?
before AV-plant TAG
'before (we) planted (the crops), right?'
9 J: [a][$_2$u].
yes
'yeah.'
10 R: [$_2$be-s]ia[$_3$ng],
MID-clear.field
'(we) cleared the field,'
11 D: [$_3$m][$_4$:][$_5$:].
hmm
'hmm.'
12 R: [$_4$kah n-(t)anam],
FUT AV-plant
'(we were) going to plant (crops),'
13 [$_5$ta]nye-ka=nye agi,
PV.ask-APPL=3 again
'he asked again,'
14 nga bapang Gitah rasan=(ny)e.
with father G. agreement=3
'their agreement, Gitah's father (did).'
15 be-ghubah.
MID-change
'(the agreement) changed.'

(BJM01-011-01, 00:01:53–00:02:06, Rili, Jamisah, Dewi)

However, PV constructions where A is realized as a verb adjacent pronoun and in a 'by'-phrase are still possible with a first person singular referent, as in (29).

(29) Emphasizing A with a 'by'-phrase
1 aku ng-angkit barut,
1SG AV-carry clothes
'I hang up my clothes (to dry),'

2 pas ku=katup duaghe pen-(t)anak-an tu.
then 1SG=PV.close door NMLZ-cook.rice-NMLZ DIST.DEM
'then I close the door to the kitchen.'
3 galak singgah die,
want visit 3
'she always visits,'
4 amu aku duduk di situ,
if 1SG sit LOC there,
'if I sit there,'
5 ku=katup liku duaghe.
1SG=PV.close by.1SG door
'I myself close the door.'

(BJM01-015, 00:25:50–00:25:58, Rumsiah)

In this case and others like it, the motivation for the use of the 'by'-phrase is different; it appears that *liku* 'by me' is used to emphasize A. In the example above, Rumsiah is telling Sawia about a neighbor that she is trying to avoid but who despite her best efforts always stops by her house. Thus, in line 5, she emphasizes the fact that she closes the door so her neighbor will not come over.

Based on the description of PV constructions in this section, there is not clearcut evidence for 'demotion' of A in a single recognizable passive construction wherein the expected demoted form (i.e., a 'by'-phrase) and demoted function (e.g., an ephemeral, non-topical A) consistently overlap in the same construction. There are certainly examples where the demoted form and function overlap, such as (20) and (23b). However, there are a number of examples where the construction has the form of a passive with A in a 'by'-phrase, but functionally does not appear to share any properties with A in passives. These sorts of PV constructions are exemplified in (22), (24), and (25). It is worth noting that 'by'-phrases serve an important function in PV constructions as NPs are essentially only available in the primary argument position or in a 'by'-phrase. Thus, if A is to be expressed as an NP in a PV construction, it must be expressed in a 'by'-phrase. From the description above, the utility of expressing an NP A in a 'by'-phrase is clear for antitopicalization constructions, but for other constructions it is less clear, although its usefulness goes beyond 'backgrounding' or 'defocusing' as has been claimed for passive constructions. Even still, there are also a number of constructions that use a 'by'-phrase but not an NP. While some constructions with a second person pronoun in 'by'-phrases may, in fact, be related to politeness in passive constructions, there is no explanation for the use of other pronouns in 'by'-phrases. Finally, PV constructions with a verb adjacent noun somehow more consistently take on passive-like functions, even though they are not formally sim-

ilar to prototypical passive constructions with 'by'-phrases. In the end, PV constructions with A have a variety of forms and functions, and while some have both the form and functions of prototypical passives, the vast majority do not.

3.3 PV without A

The majority of PV constructions occur without an A argument, as Figure 1 shows. For verbs prefixed with *di-*, this may not be surprising, and even expected, if they are indeed considered passive constructions. In fact, while the prefix *di-* has a relatively equal distribution across different realizations of A (i.e., third person pronouns, nouns, 'by'-phrases), there is a slight preference for the prefix *di-* in PV constructions that lack A (60% of all PV constructions without A), pointing to a correlation between *di-* and an unrealized A in PV constructions. At the same time, the fact that a full 40% of PV constructions without A occurs with unmarked verbs is indeed surprising. Obligatory A arguments are a hallmark of descriptions of unmarked PV constructions (i.e. Object Voice) in Malay and other western Indonesian languages (see e.g. Kroeger 2014). As the analysis of unmarked PV constructions below demonstrates, there are apparently no obligatory arguments in Besemah. Furthermore, the unmarked PV construction also has overlapping properties with apparent passive constructions, but does not fully align with passives as the verbal predicate is unmarked, making them similar to the so-called 'uncoded passives' (Arka & Kosmas 2005; Zúñiga & Kittilä 2019).

These constructions bring us back to basic questions raised at the outset of this chapter about the nature of transitive PV constructions and apparent passive constructions. In particular, what is the status of this unrealized A? Does it behave differently than other unrealized arguments? As the remainder of this section demonstrates, PV constructions without A are not uniform. A receives different interpretations, as with other instances of allusive reference, each within its local context. I discuss a range of constructions, some of which are common and expected (e.g. imperatives with an implicit second person referent) and others which are not (e.g. constructions where it is not possible to 'reconstruct' a grammatical form of A), showing that the A in PV is consistent with allusive reference found elsewhere in Besemah (Section 2.3) and in languages like colloquial Indonesian (Djenar et al. 2018; Ewing 2024). This description differs from that of Acehnese by Legate (2012) where apparent agentless passives receive an existential interpretation.

PV constructions without A comprise a diverse set. While it is not necessarily common, there are examples of PV constructions without A that resemble prototypical agentless passive constructions, in which A is 'unknown' and receives an existential interpretation (e.g. 'someone') (Zúñiga & Kittilä 2019; Legate 2021).

Examples of these types of PV constructions are found in (30), in which the verb is prefixed with *di-* in line 1 but unmarked in line 3.

(30) Passive-like interpretation in PV without A
1 J: jauh nak [di-ja][$_2$lan-i].
far want PV-walk-APPL.LOC
'(the field) is far to walk to.'
2 R: [###]—
3 [$_2$au nak jal]ani jauh anye.
yes want PV.walk-APPL.LOC far but
'yes, (it) is far to walk to.'
(BJM01-011-01, 00:01:40–00:01:43, Jamisah, Rili)

A common use is as an imperative, and while the verb is commonly unmarked in these constructions, as in (31), there are examples when the verb is prefixed with *di-*, especially in a negative imperative construction, as in line 3 of (32). Note also that imperatives are not restricted to PV but can also occur with AV, as in line 2 of (32).

(31) Unmarked PV imperative
jual-ka kudai .. perincit tu.
PV.sell-APPL first white-eye.bird DEM.DIST
'sell (your) white-eye bird.' (BJM01-130-01, 00:57:28–00:57:30, Celit)

(32) Negative imperative with PV prefixed with *di-*
1 pacak jiku.
can QUOT.1SG
'(I) can, I said'
2 .. m-buat=lah arani
AV-make=FOC QUOT
'make (it), he said.'
3 anye ame dimak jangan @ di-ulang-i @.
but TOP not.pleasant NEG.IMP PV-repeat-APPL
'but if (it's) not delicious, don't repeat (it).'
(BJM01-011-01, 00:04:56–00:05:00, Rili)

There are a number of other constructions where A is the addressee, such as a question, as in (33). In this example, someone passes by off-camera and Ujang jokingly asks him if he is going to watch the video that they are currently recording.

(33) Question for addressee
tuntun petang kele eh?
PV.watch evening later TAG
'(you'll) watch later this afternoon, right?'
(BJM01-086-01, 00:23:41–00:23:43, Ujang)

The examples thus far have demonstrated contexts in which A is more or less expected to be unrealized: apparent passive constructions with unknown or existential A and addressees in imperatives and questions. While this may be true, there is no reason to believe that these unrealized arguments are not also allusive reference in the same way that the fluid and indeterminate referents are allusive in Section 2.3. In fact, allusive reference involving the addressee or an unknown or existential A in a passive construction may be among the most straightforward for the participants and the analyst to recognize in a wide variety of contexts.

In other PV constructions without A and an unmarked verb, it is not always so straightforward, especially for the analyst. On the surface, A appears to be understood to be the speaker in line 2 of (34). In this example, Burhimin is talking to Aripin and Damsi about planting cocoa in their fields. In line 1, Burhimin states that he has a field with one area, which contrasts with the immediately preceding discourse in which Aripin talks about his field with multiple areas. In line 2, he states that half (of the field) was planted (with cocoa).

(34) Allusive first person referent in unmarked PV
1 aku kebun gi se-bidang,
1SG field only one-area
'I only have a field with one area,'
(1.7)
2 tanam-i se-piak,
PV.plant-APPL.LOC one-half
'(I) planted half,'
(BJM01-010-01, 00:15:35–00:15:40, Burhimin)

Based on the context within the conversation and knowledge of farming practices more generally, it is most likely Burhimin himself who planted the cocoa in his field in line 2. Thus, the inference of the participants is likely the speaker and the allusive reference of A. However, without any other evidence in this conversation, it is possible that he is alluding not just to himself but others who planted with him as well. Whatever inference the participants understood, the conversation proceeded without issue (Ewing 2024).

The structure of the PV construction in line 2 is particularly interesting given previous descriptions of Malay and Indonesian. If it is analyzed as an unmarked

PV construction similar to an Object Voice construction with a first or second person A in other Malay languages, it shows that A arguments are in fact not obligatory in Besemah. If, on the other hand, it is analyzed as a passive, in which case the free translation in line 2 would be 'half was planted', then Besemah provides evidence for an unmarked agentless passive in western Indonesia, which to my knowledge has yet to be attested. In fact, this example shows quite clearly that a combination of minimal structures and allusive reference in Besemah means both sorts of analysis can be true.

However, there are also PV constructions that are prefixed with *di-* where A alludes to the speaker and other speech act participants. In such cases, it is not grammatically possible to 'reconstruct' a first or second person A argument. Consider the example in (35).

(35) Allusive first person A with *di-* prefixed verb

1 H: dide di-bukak hilim,
NEG PV-open helmet
'(I/we) didn't take off (my/our) helmet(s),'

2 di-singit-ka b[aih],
PV-open-CAUS just
'(I/we) just opened (my/our mask/s) a little.'

3 I: [di]de m-buka[$_2$:k]
NEG AV-open
'(we) didn't take off (our helmets).'

4 H: [$_2$Magui] .. diam di belakang.
M. quiet LOC back
'Magui was quiet on the back.'

(0.6)

5 la ku=entak ngaghi base Indonesia lughus tu tadi.
PFV 1SG=PV.request with language Indonesia direct DEM.DIST earlier
'I answered right away in Indonesian.'

(BJM01-004-01, 00:49:28–00:49:35, Hendi, Iril)

In the excerpt in (35), Hendi is telling the other participants how he and Iril (whose nickname is Magui, mentioned in line 4) managed to get access to a good location to watch the race by playing up their importance and using Indonesian with the security guard instead of Besemah. In the first two lines, Hendi uses *di-*marked verbs without A to express that when they pulled up to the entrance with a security guard, he and potentially also Iril didn't take off his/their helmet(s), and he and potentially also Iril just opened his/their mask(s) a little. Iril, in line 4, reinfornces what Hendi is saying using an AV-marked verb, again without any

arguments. In the last two lines explicit arguments are employed by Hendi to explain who was quiet (line 5) and who talked to the security guard (line 6).

In this example, the allusive A referent in lines 1–3 again includes the speaker, but it is unclear whether the referent also includes another participant in the conversation (i.e. whether A is first person singular or plural).[8] Based on the examples of allusive reference above, this indeterminacy and fluidity of reference is maybe not surprising. In the treatment of apparent passive constructions, what is noteworthy is the fact that it is not possible to include a first person A argument when the verb is prefixed with *di-*, not even in a 'by'-phrase. In this case, it is not possible to simply 'reconstruct' a referent. If arguments were simply 'elided' or 'omitted', PV constructions like (35) would not be possible.

In addition to PV constructions involving speech act participants, there are also PV constructions in which the unrealized A is inferred to be a third person referent. Such reference is often specific for participants, even if it is not immediately clear to analysts. In the example in (36), Rafles comments that Teris is angry in line 1 and in line 3 explains that his motorcycle engine had just been rebuilt using an AV-marked verb and an allusive reference to the A argument. In line 4, Piter concurs with Rafles using a semantically similar verb, but now in PV and with an allusive A. While there is not direct evidence from the interaction, Teris is known to work on his own motorcycles, so he is the obvious A referent in lines 3 and 4.

(36) PV construction with specific A

1 R: mang Teris la imusi.
uncle T. PFV angry
'Teris was angry.'

2 P: mm imusi.
mhmm angry
'mhmm, (he) was angry.'

3 R: mutur=(ny)e empai udim m-bungkar mesin,
motorcycle=3 REC.PST finish AV-dismantle engine
'his motorcycle, (he) just finished rebuilding the engine,'

4 P: empai udim di-limir.
REC.PST finish PV-service
'(he) just serviced (the engine).'

(BJM01-004-01, 00:54:52–00:54:57, Rafles, Piter)

8. In providing free translations, Hendi used first person singular pronouns in the free translations in lines 1 and 2, but first person plural for Iril in line 3.

In this section, I have presented a range of PV constructions without A, providing examples where allusive reference to A varies. In this way, PV constructions without A in Besemah differ from apparent agentless passive constructions in Acehnese, which only receive an existential interpretation (Legate 2012). At the same time, it is not possible to simply 'reconstruct' omitted A arguments as one would expect in a 'pro-drop' scenario. Thus, PV constructions without A are best understood as any other construction in Besemah that allows for allusive reference. This analysis, in turn, leads to a new characterization of PV constructions, which do not require A arguments when the verb is unmarked. A surprising finding given how widespread obligatory A arguments are in unmarked PV (i.e. Object Voice).

4. Conclusion

In this chapter, I describe PV constructions in Besemah, showing how they are diverse and vary along two dimensions: (i) marking on the verb and (ii) the realization of A. On the surface, it is tempting to analyze a subset of these PV constructions as passive constructions (i.e. where the verb is marked with a prefix *di-* and A is either unrealized or occurs in a 'by'-phrase). However, when investigating language-in-use it is untenable to distinguish transitive PV constructions from passive constructions. Three key findings emerge from this study. First, A arguments are not obligatory in unmarked PV constructions. Second, 'by'-phrases serve multiple functions and it is unclear whether A demotion is in fact one of these functions. Third, PV constructions without A arguments exhibit the same patterns of allusive reference found elsewhere in Besemah.

Funding

This research is based upon work supported by the National Science Foundation under Grant BCS–1911641. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The Besemah corpus was compiled with funding support from a Fulbright scholarship, Blakemore Freeman Fellowship Language Grant, and Fulbright-Hays Doctoral Dissertation Research Abroad scholarship. Permission to conduct research on Besemah at that time was provided from the Ministry of Research and Technology in Indonesia.

Acknowledgements

I would like to thank Sonja Riesberg, Michael Ewing, and two anonymous reviewers for valuable feedback on this chapter. I would like to acknowledge my research counterpart in Indonesia, Yanti (Atma Jaya Catholic University of Indonesia), and to the National Research and Innovation Agency (BRIN) and the Ministry of Research and Technology in Indonesia for allowing me to carry out this research. I am particularly grateful to Hendi Feriza who provided his expertise in the transcription, translation, and analysis of the conversations and to all of the Besemah people who participated in the conversations in the corpus and supported this research more generally, most especially, Asfan Fikri Sanaf, Kencana Dewi, and Sarkani.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following abbreviations are used:

AV	A-voice	PLUR	pluractional
EXCLAM	exclamative	PV	P-voice
FILL	filler	REC	recent
MID	middle voice	TAG	tag question
N.LI	light noun	TRUN	truncated

References

Adelaar, Alexander. 2005. "Structural Diversity in the Malayic Subgroup." In *The Austronesian Languages of Asia and Madagascar*, ed. by Alexander Adelaar & Nicholas P. Himmelmann, 202–226. New York: Routledge.

Arka, I Wayan. 2019. "The Core-Oblique Distinction in Some Austronesian Languages of Indonesia and Beyond." *Linguistik Indonesia* 35 (2): 101–144.

Arka, I Wayan, and Jeladu Kosmas. 2005. "Passive Without Passive Morphology." In *The Many Faces of Austronesian Voice Systems: Some New Empirical Studies*, ed. by I Wayan Arka, and Malcolm Ross, 87–117. Canberra: Pacific Linguistics.

Arka, I Wayan, and Christopher D. Manning. 2008. "Voice and Grammatical Relations in Indonesian: A New Perspective." In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Simon Musgrave, and Peter K. Austin, 45–69. Stanford: Center for the Study of Language.

Chafe, Wallace L. 1976. "Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View." In *Subject and Topic*, ed. by Charles N. Li, 25–55. New York: Academic Press.

Chen, Victoria, and Bradley McDonnell. 2019. "Western Austronesian Voice." *Annual Review of Linguistics* 5 (1): 173–195.

Cole, Peter, Gabriella Hermon, and Yanti. 2008. "Voice in Malay/Indonesian." *Lingua* 118 (10): 1500–1553.

Cumming, Susanna. 1991. *Functional Change: The Case of Malay Constituent Order*. Berlin: Mouton de Gruyter.

Djenar, Dwi Noverini, Michael Ewing, and Howard Manns. 2018. *Style and Intersubjectivity in Youth Interaction*. Berlin, Boston: De Gruyter.

Donohue, Mark. 2007. "Variation in Voice in Indonesian/Malay: Historical and Synchronic Perspectives." In *Diversity in Language: Perspectives and Implications*, ed. by Yoshiko Matsumoto, David Oshima, Orrin Robinson, and Peter Sells, 71–129. Stanford, CA: CSLI Publications.

Du Bois, John W., Susanna Cumming, Stephan Schuetze-Coburn, and Danae Paolino. 1992. *Discourse Transcription*. Santa Barbara: Department of Linguistics, University of California.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. "Outline of Discourse Transcription." In *Talking Data: Transcription and Coding in Discourse Research*, ed. by Jane Anne Edwards, and Martin D. Lampert, 45–89. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Durie, Mark. 1988. "The So-Called Passive of Acehnese." *Language* 64 (1): 104.

Eades, Domenyk. 2005. *A Grammar of Gayo: A Language of Aceh, Sumatra* Vol. 567. Canberra: Pacific Linguistics.

Ewing, Michael C. 2014. "Motivations for First and Second Person Subject Expression and Ellipsis in Javanese Conversation." *Journal of Pragmatics* 63. 48–62.

Ewing, Michael C. 2019. "The Predicate as a Locus of Grammar and Interaction in Colloquial Indonesian." *Studies in Language* 43 (2): 402–443.

Ewing, Michael C. 2024. "Elusive Referentiality and Allusive Reference in Indonesian Conversation." In *Pragmatics & Beyond New Series*, Vol. 344, ed. by Michael C. Ewing, and Ritva Laury, 11–34. Amsterdam: John Benjamins Publishing Company.

Ewing, Michael C. 2025. "The Atypicality of Predicates with Two Explicit Arguments in Indonesian Conversation." *Languages* 10 (2): 28.

Fox, Barbara A. 1994. "Contextualization, Indexicality, and the Distributed Nature of Grammar." *Language Sciences* 16 (1): 1–37.

Givón, Talmy. 1983. *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Amsterdam: John Benjamins.

Givón, Talmy (ed). 1994. *Voice and Inversion*. Amsterdam: John Benjamins.

Goffman, Erving. 1983. "Felicity's Condition." *American Journal of Sociology* 89 (1): 1–53.

Haspelmath, Martin. 1990. "The Grammaticization of Passive Morphology." *Studies in Language* 14 (1): 25–72.

Himmelmann, Nikolaus P. 2005. "The Austronesian Languages of Asia and Madagascar: Typological Characteristics." In *The Austronesian languages of Asia and Madagascar*, ed. by Alexander Adelaar, and Nikolaus P. Himmelmann, 110–181. New York: Routledge.

Jeoung, Helen. 2017. "On the Number of Voices in Madurese." *Journal of the Southeast Asian Linguistics Society* 10 (1): 16–35. http://hdl.handle.net/10524/52395

Kazenin, Konstantin I. 2001. "The Passive Voice." In *Language Typology and Language Universals: An International Handbook*, Vol. 2, ed. by Martin Haspelmath, 899–915. Berlin: Walter de Gruyter.

Keenan, Edward L., and Matthew S. Dryer. 2007. "Passive in the World's Languages." In *Language Typology and Syntactic Description: Clause Structure*, 2nd ed., Vol. 1, ed. by Timothy Shopen, 325–361. Cambridge: Cambridge University Press.

Kroeger, Paul. 2014. "Passive Agents in Malay: The Binding Properties and Discourse Functions of Agentive =nya." In *Current Trends in Malay Linguistics*, ed. by Siaw-Fong Chung, and Hiroki Nomoto, 5–29.

Kroeger, Paul, and Sonja Riesberg. 2024. "Voice and Transitivity." In *The Oxford Guide to the Malayo-Polynesian Languages of Southeast Asia*, ed. by Alexander Adelaar, and Antoinette Schapper, 795–824. Oxford: Oxford University Press.

Laskowske, Douglas. 2016. Voice in Bugis: An RRG Perspective. Grand Forks: University of North Dakota dissertation.

Lawler, John M. 1988. "On the Questions of Acehnese 'Passive'." *Language* 64 (1): 114–117.

Legate, Julie Anne. 2012. "Subjects in Acehnese and the Nature of the Passive." *Language* 88 (3): 495–525.

Legate, Julie Anne. 2021. "Noncanonical Passives: A Typology of Voices in an Impoverished Universal Grammar." *Annual Review of Linguistics* 7 (1): 157–176.

McDonnell, Bradley. 2008-present. *Besemah (BJM01)*. PARADISEC.

McDonnell, Bradley. 2016. Symmetrical Voice Constructions in Besemah: A Usage-based Approach. Doctoral Dissertation: University of California, Santa Barbara.

McDonnell, Bradley. 2022. "Universal Quantifiers, Focus, and Grammatical Relations in Besemah." *Studies in Language* 47 (2): 422–462.

McDonnell, Bradley, and Victoria Chen. 2022. "The Evolution of Syntax in Western Austronesian." In *The Routledge Handbook of Asian Linguistics*, ed. by Chris Shei, and Saihong Li, 11–32. London; New York: Routledge.

McDonnell, Bradley, Jiang Wu, Timothy McKinnon, and Alexander Adelaar. 2024. "The Malayic languages." In *The Oxford Guide to the Malayo-Polynesian Languages of Southeast Asia*, ed. by Alexander Adelaar, and Antoinette Schapper, 431–454. Oxford/New York: Oxford University Press.

McDowell, Jonathan, and Karl Anderbeck. 2020. *The Malay Lects of Southern Sumatra*. Honolulu: University of Hawai'i Press.

Nariyama, Shigeko. 2003. *Ellipsis and Reference Tracking in Japanese* Vol. 66. Amsterdam: John Benjamins.

Oh, Sun-Young. 2007. "Overt Reference to Speaker and Recipient in Korean." *Discourse Studies* 9 (4): 462–492.

Patrianto, Hero, and Victoria Chen. 2023. "Two Sides to the Same Coin: Reappraising Indonesian-Type 'Passive' and Object Voice in Javanese." In *Proceedings of TripleAFLA: 9th TripleA Workshop for Semantic Fieldworkers, 29th Annual Meeting of the Austronesian Formal Linguistics Association*, ed. by Vera Hohaus, Jens Hopperdietzel, and Siena Weingartz, 59–74. London/Ontario: Western University.

Riesberg, Sonja. 2014a. *Symmetrical Voice and Linking in Western Austronesian Languages*. Berlin: De Gruyter.

Riesberg, Sonja. 2014b. "Passive Actors Are Not Adjuncts — Consequences for the Distinction Between Symmetrical and Asymmetrical Voice Alternations." In *Argument Realisations and Related Constructions in Austronesian Languages: Papers from 12-ICAL*, Vol. 2, ed. by I Wayan Arka, and N. L. K. Mas Indrawati, 281–302. Canberra: Asia-Pacific Linguistics.

Shibatani, Masayoshi. 1985. "Passives and Related Constructions: A Prototype Analysis." *Language* 61 (4): 821–848.

Zúñiga, Fernando, and Seppo Kittilä. 2019. *Grammatical Voice*. Cambridge: Cambridge University Press.

CHAPTER 22

The acquisition of symmetrical voice systems

Comparative notes on Tagalog, Indonesian, and Totoli

Maria Bardají, Rowena Garcia, Christoph Bracks & Evan Kidd
Universitat de Barcelona | University of the Philippines Manila | University of Frankfurt | Australian National University

Language documentation has the broad aim of building a representative evidence base across typologically diverse languages and contexts of use. One such context that is gaining increased attention is language acquisition, owing to its importance in language maintenance in an era of significant language endangerment and loss. In this chapter we discuss the acquisition of a typologically important feature of Austronesian languages — symmetrical voice — to which Nikolaus Himmelmann has made distinguished contributions. We consider three languages that span the commonly made divide between Philippine-type and Indonesian-type languages: Tagalog, Jakarta Indonesian, and Totoli. We first compare the previous literature on the acquisition of Tagalog and Indonesian, for which there is significant data. We then consider how Totoli differs from Philippine- and Indonesian-type languages, and, based on a preliminary corpus observation, make predictions on the acquisition process of the Totoli voice system.

Keywords: symmetrical voice, child language documentation, child-directed speech, Tagalog, Totoli, Indonesian, western Austronesian

1. Introduction

Owing to the "documentalist turn" in the field of language documentation — spearheaded by Himmelmann (1998) — more and more data on small and under-described languages have become available (see Seifart et al. 2018). These data support richer linguistic descriptions, more representative typological studies,

https://doi.org/10.1075/slcs.240.22bar
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

and ultimately more robust theoretical insights. Some three decades later, other subfields of linguistics are currently reckoning with a distinct lack of data on diverse languages. One such field is child language acquisition: despite a strong history of cross-linguistic research, the field remains heavily skewed towards studies of English and other major European languages (Kidd & Garcia 2022; Passmore et al. 2025). There are, however, recent initiatives aimed at countering this bias, such as the Acquisition Sketch Project (Hellwig et al. 2023), which offers guidance in the collection of acquisition data across many languages, especially in contexts of language endangerment. This project aligns with Himmelmann's (2006: 2) guidelines of "casting the net as wide as possible" in a documentation collection. That is, on the one hand, efforts in the collection of language by children and directed to children work in favour of a thorough documentation of a language that embraces as many linguistic practices as possible. On the other hand, this benefits the discipline of language acquisition research by enabling the study of the acquisition of typologically diverse structures and language learning contexts.

It is precisely at the intersection of language acquisition research, language documentation, and linguistic typology that the present paper is situated. Following the motto of this volume "from observable linguistic behaviour to typological generalizations", this paper investigates how different types of symmetrical voice systems are acquired. This system found in many western Austronesian languages is of interest to typology and language acquisition for the following reasons. Unlike in asymmetrical systems such as the active-passive alternation, symmetrical voice languages have more than one basic transitive construction. That is, languages exhibit at least two constructions, called here actor voice (AV) and undergoer voice (UV),[1] which both have two core arguments (see Riesberg 2014 for a full account of symmetrical voice). The difference between the two voices pertains to the thematic role of the subject argument. This is exemplified in (1) by Madurese. In an AV construction like (1a), the subject argument is an agent, whereas in an UV construction such as (1b), the subject has the thematic role of undergoer. Crucially, unlike a passive clause, the UV is also a transitive construction: both *Bibbi'* 'auntie' and *na'kana'* 'the children' are core arguments and, consequently, the agent argument is not demoted to an oblique.

1. The term undergoer will be used throughout this paper as a cover term for various undergoer-like semantic roles (e.g. theme, patient, recipient). Some languages, such as Tagalog, have distinct voices corresponding to specific types of undergoers (e.g. patient voice, locative voice). If a particular voice is meant, we will refer to it by its specific name, otherwise we will use *undergoer*.

(1) Madurese (Davies 2010: 250)

a. *Bibbi' nenggu na'-kana'.*
aunt AV.watch RDP-child
'Auntie watched the children.'

b. *Na'-kana' e-tenggu Bibbi'.*
RDP-child UV-watch aunt
'Auntie watched the children.'

From an acquisition perspective, symmetrical voice languages present complex challenges to the child language learner, for whom a fundamental early task in acquisition is to acquire core grammatical categories and how they encode meaning in utterances. In symmetrical voice languages, children learn more than one transitive construction, of which at least one does not link the agent argument to the subject function by default. In fact, in many Austronesian languages the more preferred mapping is between the patient and the grammatical subject. In addition, in terms of word order, western Austronesian languages do not necessarily place the agent in first position (see Example 1b above). Thus, in terms of thematic role assignment, symmetrical voice languages might be challenging for language acquisition because heuristics such as interpreting the first noun that they hear as the agent are suboptimal and cannot be relied on like they might be in other languages (see Pine et al. 2025), and hypothesised innate linking rules mapping the agent to the subject role do not work (Pinker 1984, 1989). Instead, children must pay more attention to verbal morphology (and in addition either noun morphology or word order) in order to properly assign thematic roles to the arguments of the clause.

There are, however, structural differences among symmetrical voice languages, which suggests that acquisition trajectories may differ across different languages. A common typological distinction is drawn between Philippine-type languages (e.g. Tagalog) and Indonesian-type languages (e.g. Indonesian). However, certain languages of North and Central Sulawesi (such as Totoli) appear to fall somewhere in between, potentially constituting a third type. While these typological distinctions have been explored in terms of verbal morphology in adult language, their implications for language acquisition remain underinvestigated.

In this context, this paper focuses on the acquisition of three languages — Tagalog, Jakarta Indonesian, and Totoli — each representing one of the three types of symmetrical voice systems defined in detail in Section 2 below. To begin with, one of the main concerns is the type of data available for these languages, which differ in many regards: there is experimental and corpus data for Tagalog and Jakarta Indonesian, which are major languages with millions of speakers, and only a documentation corpus for Totoli, an endangered language with at most 5.000 speakers. Issues related to the data are discussed in Section 3.1.

The second question addressed in this paper is whether there are differences in the acquisition of the three languages that are influenced by the previously mentioned typological distinction. To this end, we review the input received by children (Section 3.2), the production of voice-marked predicates (Section 3.3), strategies for thematic role assignment (Section 3.4) and the relative word order of agent and patient arguments (Section 3.5). Observations on Tagalog and Jakarta Indonesian come from a literature review, while all claims on Totoli are predictions based on a preliminary corpus observation and on our knowledge of how this language differs from the Philippine and the Indonesian type. Finally, the conclusion in Section 4 addresses the relevance of the differences and commonalities found among the three languages. On the one hand, they illustrate that data on language acquisition can be used for typological characterisations. On the other hand, they are relevant for acquisition theories, especially with regard to the fact that a construction in which the agent is not the subject argument is acquired from an early age and apparently with the same ease that a construction which does link the agent role with the subject function.

2. Typology of western Austronesian symmetrical voice

While the brief characterisation of symmetrical voice given above holds for the vast majority of symmetrical voice languages, there are structural differences among them with respect to several parameters involved in the alternation. A major distinction is commonly made between Philippine-type languages and Indonesian-type languages (Wolff 1996; Arka 2003; Himmelmann 2002b, 2005a; Chen & McDonnell 2019). This divide is made on a typological rather than phylogenetic basis. The distinction concerns mostly the verbal system of the languages, but differences are also found in other areas. In terms of areal distribution, the distinction between Philippine and Indonesian languages roughly coincides with their geographical distribution except for the languages of Borneo and North and Central Sulawesi. Languages of North Borneo are usually included in the group of Philippine languages, whereas languages of South Borneo belong to the group of Indonesian languages. North and Central Sulawesi languages seem to be typologically somewhere between Philippine and Indonesian languages and we treat them here as a separate group.

In this section, we describe the characteristic features of the voice system in each type, without entering a discussion of the minimal set of features that need to be shown by each language in order to be considered of one type or the other. In addition to the Philippine type (Section 2.1) and the Indonesian type (2.2), we distinguish the group of North and Central Sulawesi languages (2.3), which repre-

sent an area of transition. Finally, in Section 2.4, we discuss further areas of grammar where the three groups differ.

2.1 The voice system of Philippine-type languages

Philippine-type languages exhibit a voice system which is probably closest to that of Proto-Austronesian (Wolff 1996: 16). Languages of this group have, besides an actor voice, at least two formally and semantically different undergoer voices (Himmelmann 2005a: 113). This is exemplified in (2) by Tagalog. Besides the actor voice illustrated in (2a), Tagalog has three different undergoer voices: a patient voice (PV) marked with the suffix *-in*, in which the subject is a patient (2b); a conveyance voice (CV) marked by *i-*, in which the subject has the thematic role of a (displaced) theme (2c); and a locative voice (LV) marked by *-an,* in which the subject expresses either a location or the recipient or goal of the action (2d).

(2) Tagalog

a. *h<um>anap na ng bahay ang mga bata*
<AV>search already GEN house SPEC PL child
'The children looked for houses/a house.' (Himmelmann 2005b: 356)

b. *patay-in natin ito-ng dalawa-ng Hapon*
dead-PV 1PLi.GEN PRX-LK two-LK Japan
'Let's kill these two Japanese.' (Himmelmann 2005b: 367)

c. *iniabot ng manggagamot sa sundalo ang itlog*
i<in>-abot ng manggagamot sa sundalo ang itlog
CV<RLS>-reach GEN doctor LOC soldier SPEC egg
'the physician handed the egg to the soldier' (Himmelmann 2008: 265)

d. *pinuntahan na naman nila ang bata*
p<in>unta-an na naman nila ang bata
<RLS>direction-LV now also 3PL.GEN SPEC child
'They went to the child.' (Himmelmann 2005b: 366)

Regarding dynamic verbs,[2] Philippine-type languages show three main features. First, all verbs are obligatorily marked for voice, even intransitive verbs, which are marked for actor voice (Chen & McDonnell 2019: 176). Second, verbs show tense/mood and aspect distinctions (Wolff 1996: 16): voice alternations usually come in two moods (realis and non-realis) and aspectuality is generally marked by means of reduplication (cf. the full paradigm in Table 1). In addition, voice-

2. Most western Austronesian languages make a morphological distinction between dynamic and stative events. All claims throughout this paper only concern dynamic verbs. Stative verbs may also take part in the voice alternation, but they differ in several regards and are therefore excluded from the characterisation.

marked forms do not combine with applicative morphology: it is the voice alternation which allows micro-roles to occupy the subject position (see for instance, 2d above) (Arka 2003).

Table 1. Tagalog voice-mood paradigm based on Himmelmann (2005b: 363)

	AV	**PV**	**LV**	**CV**
non-realis/ perfective	b<um>ili	bilh-in	bilh-an	i-bili
non-realis/ imperfective	bi~bili	bi~bilh-in	bi~bilh-an	i-bi~bili
realis/ perfective	b<um>ili	b<in>ili	b<in>ilh-an	i-b<in>ili
realis/ imperfective	b<um>i~bili	b<in>i~bili	b<in>i~bilh-an	i-b<in>i~bili

Philippine-type languages are also distinguished from Indonesian-type languages by characteristic argument marking patterns. According to Himmelmann (2005a: 113), these languages have at least one non-local phrase-marking clitic used to mark nominal expressions. As shown in (2), it is common for Philippine languages to have a marker for the subject argument (*ang* in the examples above), a non-subject phrase marker (*ng* above) and another marker for locative or oblique arguments (*sa* in (2c)). These markers as well as pronominal forms usually come in two paradigms: one for personal names and one for all other expressions. A further characteristic noted by Himmelmann (2005: 113) is that Philippine-type languages have pronominal second position clitics, such as *natin* in (2b).

2.2 The voice system of Indonesian-type languages

The Indonesian-type alternation is considered a less complex system that has developed from the Philippine-type system (Wolff 1996: 16). These languages show a two-way distinction between an actor voice and an undergoer voice, as shown in (1) above. In terms of verbal marking, actor voice verbs are usually marked, typically by a nasal prefix (but not necessarily, as is shown below for Jakarta Indonesian), whereas undergoer voice predicates may either be marked by a prefix or have no marking at all (Chen & McDonnell 2019: 178). Unlike in Philippine-type languages, there is variation in the marking of dynamic intransitive verbs: they may come in the actor voice, but there are other intransitive prefixes (e.g. *ber-* in Indonesian) and even unmarked forms. Verbs have also lost the TAM distinctions seen in the Philippine languages above (see Wolff 1996 for the diachronic path). A further difference regarding the verb is that in Indonesian-type languages voice marking can combine with applicative morphology, mostly with a valency-increasing function (Himmelman & Riesberg 2013). An example is given in (3).

(3) Indonesian (Arka 2003: 3–4)
a. *Kursi itu dia Ø-duduk-i*
chair DIST 3SG UV-sit-APPL
'The chair, (s)he sat on it.'
b. *Batu itu mereka Ø-lempar-kan ke rumah itu*
stone DIST 3PL UV-throw-APPL to house DIST
'The stones, they threw them to the house.'

Nominal expressions are generally not preceded by phrase markers (exceptions exist, e.g. Tukang Besi, Arka 2003: 2). As for pronouns, in Indonesian-type languages, some persons have been integrated into the voice system as pronominal proclitics or prefixes (Wolff 1996: 28; Himmelmann 2005a: 169–170). Consider the Indonesian example in (4).

(4) Indonesian (Himmelmann 2005a: 136)
orang itu ku-lihat
person DIST 1SG-see
'I saw that person.'

Finally, a major characteristic of Indonesian-type languages is that they have developed a true passive construction with a demoted agent, which is either unexpressed or demoted to an oblique phrase. Consider the example in (5).

(5) Balinese high register (Arka & Simpson 2008: 98)
Nyoman sampun ka-rereh (antuk ida)
PN PRF PASS-search by 3SG
'Nyoman has been searched for (by him/her).'

Himmelmann (2002a, 2002b, among others), however, notes that the languages generally subsumed under the term Indonesian-type languages do not constitute a homogeneous group in the way that Philippine-type languages do. Thus, a systematic examination of the group of non-Philippine-type languages remains a pending task — one that is necessary to determine which languages conform to the characteristics described in this subsection and which display divergent patterns. Among the latter, languages of North and Central Sulawesi have been mentioned in the literature — especially by Wolff (1996), Himmelmann (2002b), and Arka (2003) — as not fitting neatly into the features of Indonesian-type languages (nor of Philippine-type languages). Given that we are concerned with one of these languages — namely Totoli — in the next section we characterise this set of languages as a further typological (sub)group. It is also likely that other transition areas between the Philippine and the Indonesian types exist. Good candidates are the languages of Borneo (see for instance Hemmings 2015 on Kelabit (Northern Sarawak), which also has mixed features) but this is beyond the scope of this research.

2.3 The voice system of North and Central Sulawesi languages

Languages of North and Central Sulawesi represent an area of transition between the Philippine type and the Indonesian type, both geographically and in terms of structure, especially with regard to the verbal system.[3] These languages differ from Indonesian-type languages (and thus pattern with Philippine-type languages) in that they have not developed a true passive construction and have retained mood and aspect distinctions. However, they still differ from Philippine-type languages in several regards. First, languages from North and Central Sulawesi do not have obligatory phrase markers for nominal expressions. Second, they might have more than one undergoer voice construction, but these are not always clearly distinguishable formally and semantically. This is illustrated by the following voice-mood paradigms of Ratahan and Totoli in Tables 2–3. In the Ratahan paradigm, the same formatives are used to build the different undergoer voices and identical formations can only be distinguished if they are considered with respect to other forms in the paradigm (e.g. the realis equivalent of a form suffixed with *-an* indicates whether this is a bare undergoer voice or a locative one). The Totoli paradigm, besides showing some degree of formal syncretism among the different formatives, also shows two undergoer voices which are not distinguishable semantically: verbs are lexically determined for one of the two undergoer voices but they do not seem to differ semantically.

Table 2. Basic voice-mood paradigm in Ratahan (from Himmelmann 2002b)

	Realis	**Non-realis**
actor voice	<im> / N-	<um> / M-
undergoer voice	<in>	-an
undergoer.locative voice	<in> -an	-an
undergoer.theme voice	<in>	Ø

Some North and Central Sulawesi languages also have applicative formations. See, for instance, Example (6) from Totoli: *taip* is a transitive root, which with applicative morphology can accept a third argument (a beneficiary). However, the use of these constructions illustrates that these languages are neither clearly from the Philippine type nor from the Indonesian type, but somewhere in between. While Indonesian languages have productive applicative affixes, Philippine lan-

3. These languages are rarely treated as an additional typological subgroup as we do here. Further research is thus needed for a more precise picture of the languages that conform to this group.

Table 3. Totoli dynamic voice-mood paradigm

	Realis	**Non-realis**
actor voice	no-/nog-/non-	mo-/mog-/mon-
undergoer voice 1	ni-	Ø (bare root)
undergoer voice 2	ni- -an	-i
locative voice	ni-po-/pog-/pon- -an	po-/pog-/pon- -i

guages lack them (voice alternations are used instead). In North and Central Sulawesi languages, applicative morphology is closely intertwined with voice morphology in a way that their functions cannot be easily disentangled. For example, in (6), different applicative formations select a different role for the subject function: in (6a) it is the theme, while in (6b) it is the applied beneficiary. Regarding this topic, the reader is referred to Himmelmann and Riesberg (2013).

(6) Totoli

a. *Taipang itu kodoong taip-an=na i aku.*
mango DIST FUT slice-UV.APPL=3SG.GEN HON 1SG
'He will peel that mango for me.'

b. *Aku nipanaip Rinto taipang.*
aku ni-poN-taip Rinto taipang
1SG UV.RLS-SF-peel:APPL PN mango
'For me, Rinto peeled a mango.'

2.4 Differences beyond voice alternations: Word order and frequencies of actor and undergoer voice

Not much is known about additional differences between Philippine-type and Indonesian-type languages in other areas of grammar beyond the voice system, let alone with regard to North and Central Sulawesi languages. Himmelmann (2005a: 145) points out that Philippine-type languages have minimal inventories of primary prepositions, while Indonesian-type languages seem to have more. Moreover Philippine-type languages have pronominal second position clitics, while other languages do not. Bardají (2024) notes that the same typological division can be made based on nominalising morphology: Philippine-type languages and Indonesian-type languages show clear differences in their nominalising strategies. Furthermore, North and Central Sulawesi languages share features with both groups and, in addition, have their own nominalising devices.

Besides these, we note two potential further differences which can be relevant for the acquisition of the different symmetrical voice types. First, languages differ

with regard to word order. Philippine-type languages exhibit generally a verb-initial word order. In languages like Tagalog, the arguments following the verb tend to obey two tendencies: an agent-first and a subject-last tendency (Kroeger 1993), which can both be met in undergoer voice clauses but compete in actor voice clauses. Indonesian languages have a preference for SVO word order but this is not the only order allowed. Regarding North and Central Sulawesi languages, it remains to be investigated how homogeneous the group is, but the survey in Riesberg et al. (2019) suggests that these languages show both SVO and VOS patterns. For the whole group of western Austronesian symmetrical voice languages, Riesberg et al. (2019) also report a general tendency towards agent-first patterns.

With the data available, it seems that a second difference between Philippine-type and Indonesian-type languages could be the frequency of each voice. AV is more frequent than UV in modern Indonesian-type languages. For instance, with respect to, respectively, Besemah and Balinese, McDonnell (2016) and Pastika (1999) report a distribution of between 60% and 40% in spoken corpora. This is, however, not the case of Classical Malay, where Cumming (1991) finds more UV than AV (73% vs. 27%). In corpora of Tagalog (Adricula 2023), the distribution seems the opposite: UV is more frequent than AV. Languages of North and Central Sulawesi seem to resemble Philippine languages (or at least Tagalog). In a corpus of Totoli, Riesberg et al. (2021) find 72% of UV and 28% of AV. A similar distribution is found in a corpus of Tondano (Brickell 2014) and Pendau (although less pronounced; Quick 2005). Although distributional data in more languages are needed to confirm these tendencies, below we show that similar differences can be observed in corpora of child-directed speech, which means that the input of each voice received by children varies depending on the type of voice system.

3. The acquisition of symmetrical voice

Having spelled out the differences between the three types of symmetrical voice languages, we now look at them from the perspective of language acquisition. To this end, we focus on Tagalog (Philippine type), Jakarta Indonesian (Indonesian type) and Totoli (North- and Central-Sulawesi type). What we can infer from the previous section is that, in the acquisition of grammatical relations in transitive sentences, children are faced with three different scenarios. Tagalog children need to learn at least four different structures and they can rely on verbal morphology as well as nominal marking for thematic role assignment. Indonesian children must acquire two basic transitive constructions plus a valency-changing one (the passive) and can rely partially on verbal morphology but not on nominal marking (word order can be used instead). Totoli children must learn two different transi-

tive constructions (although with two different markings in the UV) and can map thematic roles to arguments by means of voice morphology, word order and, in some cases, the form of personal pronouns.

To begin with, we review the child language data available for each language and discuss the issues related to its collection (Section 3.1). We then analyse the input received by children with respect to symmetrical voice (Section 3.2). Finally, we address three of the most relevant aspects involved in the acquisition of a symmetrical voice system. First, the fact that children learn two different basic transitive predicates differentiated by voice-morphology (Section 3.3). Second, that this voice morphology (and in Philippine languages also nominal morphology) is used for thematic role assignment (Section 3.4). Third, we address the topic of word order with the background question of whether an agent-first preference can be observed in the acquisition of any of the three types (Section 3.5). The focus on these three aspects leaves out other grammatical parameters in which the three types differ (e.g. applicative morphology); nevertheless, comparable data for those is not yet available. All sections except the one on data follow the same structure: first, the literature on the acquisition of Tagalog and Jakarta Indonesian is reviewed and then some predictions on the acquisition of Totoli are drawn.

3.1 Child language data of western Austronesian languages

Voice systems of western Austronesian languages have received considerable attention, yet their acquisition has received less attention, even though Austronesian is the second largest language family in the world (behind Niger-Congo). Few child language corpora of western Austronesian languages exist and only a small amount of research on topics regarding language acquisition for these languages is available. The most notable exceptions being Tagalog and to a lesser extent Jakarta Indonesian. Regarding the acquisition of the Tagalog voice system, a growing number of both experimental studies as well as corpus-based studies are available (e.g. Bautista 1983; Garcia et al. 2018, 2019, 2020, 2021; Garcia & Kidd 2020; Marzan 2013; Marzan et al. 2014; Segalowitz & Galang 1978; Tanaka 2016; see also the review by Garcia & Kidd 2022:1862). Research into the acquisition of the voice system of Jakarta Indonesian has mainly revolved around the Jakarta Field Station that resulted in several studies (Gil & Tadmor 2007; Gil 2006; Hidajat 2010). Experimental data is reported by Hidajat (2010), alongside a corpus study. No research on the acquisition of Totoli has yet been conducted, as in many other western Austronesian languages.

The three languages under investigation here differ regarding their sociolinguistic setting. Tagalog and Jakarta Indonesian are major languages, each spoken by several million speakers. Totoli, on the other hand, is an endangered minor-

ity language spoken by at most 5.000 speakers in the northern part of Central Sulawesi. Besides the small number of speakers, Totoli also has a low rate of intergenerational transmission (Fatinah et al. 2024). From all Totoli-speaking villages, Totoli is only transmitted in one: namely in Gio, a village with approximately 800 inhabitants (Bracks & Himmelmann, in prep.).

This difference regarding the sociolinguistic setting is also reflected in the data available for the three languages. For Jakarta Indonesian there exists a sizeable longitudinal corpus with recordings of eight children between the age of 1:7 and 4:6 who were recorded over the course of five years (Gil & Tadmor 2007).

For Tagalog, anonymised transcriptions of a total of twenty hours of recordings of semi-structured interactions between 20 pairs of Tagalog-dominant children (2;0–3;10) and their caregivers are available at The Language Archive (https://archive.mpi.nl/: Garcia 2019). This cross-sectional data set was collected in 2019. Each pair was recorded for one hour at their homes in Caloocan City, Metro Manila, as they used the provided materials (e.g. books and toys). A more naturalistic corpus of speech from six Tagalog-English bilingual children (1;2–5;0) from Metro Manila was collected by Marzan (2013) from 2011–2013. Each child was visited and recorded at home every two months for two years, resulting in 78 30-minute recordings. However, this data set is not openly accessible. Currently, two other longitudinal corpora are being prepared and transcribed: (i) semi-structured interactions of 10 children (0;8–3;4) and their caregivers, similar to Garcia's (2019), who were recorded in their homes in Caloocan city every two weeks for two years (Garcia et al., in prep), and (ii) daylong recordings of children (0;5–6;5, mean age: 2;7) from rural Batangas, speaking the Batangas Tagalog dialect (Huber et al. 2025), who were recorded every three months for one and a half years (total of 224 day-long recordings).

Corpus data from Totoli were recorded as part of a larger language documentation initiative. To date, there are two documentation collections of the Totoli language: Leto et al. (2005–2010) and Bracks et al. (2023), which amount to more than 200 hours of recordings. The first corpus, collected by Leto et al. (2005–2010) did not specifically target the inclusion of child language and only contains one hour of interactions among elementary school children. The collection by Bracks et al. (2023) contains a subcorpus of child and child-directed language (the Totoli Child Language Corpus), which is the one on which we base the predictions of the following sections. Data were collected between 2019–2023 following the approach of the Sketch Acquisition Manual (Defina et al. 2023; Hellwig et al. 2023). It is a cross-sectional corpus with day-long recordings from 8 children between the ages of 2 and 4. The recordings amount to a total of 59 hours of recordings, although thus far about 30 minutes of speech of 4 out of the 8 children is transcribed.

3.2 The input

As mentioned in Section 2.4, corpora of symmetrical voice languages (adult-to-adult speech) indicate that (i) the different voices differ in their frequency of occurrence and (ii) this distribution might be a difference between the three types. If this is the same for child-directed speech (CDS), it means that children receive more input in one voice than in the other, which might influence their acquisition.[4] This section shows that this is the case by reviewing the voice distribution reported for CDS corpora of Tagalog (Section 3.2.1.) and Jakarta Indonesian (Section 3.2.2.). The same is predicted for Totoli in Section 3.2.3.

3.2.1 *Tagalog CDS*

From a corpus of interactions between two- to four-year-old Tagalog-speaking children and their caregivers, Garcia and Kidd (2022) found that in 9,401 child-directed utterances with verbs, 35% of verbs were uninflected for voice even though these would be inflected in adult-directed speech. An inspection of a subset of these composed of verb-initial utterances with semantically-causative verbs (where the patient is affected by the action, as *the pig* in *The cow pushed the pig,* 2,935 utterances) showed that 82% of uninflected verbs should have been in the patient voice based on real-world probability and noun marking (Garcia et al. 2021). This use of uninflected verbs was observed more often in imperative compared to non-imperative utterances (Garcia et al. 2021; Garcia & Kidd 2022).

Among the 65% of verbs which were inflected for voice, 69% were in the undergoer voice, and 31% were in the actor voice (Garcia & Kidd 2022). This, along with the finding on uninflected verbs, means that children's input usually has an undergoer voice mapping, wherein the subject is not an agent.

---

4. In this and the following sections, we focus only on the frequency of AV vs. UV structures, regardless of the root. As pointed out by a reviewer, however, specific roots might be biased towards a particular voice and thus appear more frequently in AV or UV in the input (see for instance Jap & Hsu 2025 on Indonesian adult data). As far as we know, it is yet to be investigated whether there is a correspondence between each root's voice in CDS and the one observed in children's performance in the three languages investigated. Nevertheless, several aspects suggest that it still makes sense to consider frequencies of AV and UV irrespective of the root. Most importantly, to master AV or UV, children do not only need to use the right verbal morphology but also need to assign thematic roles correctly (which in Tagalog also implies using the appropriate nominal markers). Experiments by Garcia and Kidd (2022) tested the same Tagalog roots in AV and PV configurations and children were generally better when producing sentences with a patient voice-marked verb. Finally, regarding Totoli (adult data), Riesberg et al. (2021: 36–42) show that while a few roots do show a bias towards one of the voices, the great majority of verbs in the corpus are not clearly actor- or undergoer-voice oriented.

In terms of word order, an inspection of verb-initial utterances with semantically-causative verbs showed that both agent and undergoer voices (as well as uninflected verbs) are typically agent-initial (AV: 92%, UV: 88%, uninflected: 96%; Garcia et al. 2021). However, when utterances with pronouns are excluded, agent-initial utterances are no longer as frequent in the agent voice (AV: 56%; UV: 74%, uninflected: 80%). Overall, these results indicate that most of the input that children hear are agent-initial utterances.

3.2.2 *Jakarta Indonesian CDS*

The analysis in Hidajat (2010) of a CDS corpus of Jakarta Indonesian provides data on four different verbal clauses: actor voice constructions prefixed with *N-*, clauses with unaffixed predicates but with actor voice configuration (SVO order where the subject is an agent), and two types of clauses prefixed with *di-*, namely what we call here UV constructions (i.e. transitive predicates with an overt non-demoted agent) and proper passive clauses (i.e. without an agent or with a demoted agent). If both unmarked and marked verbal forms are taken into account, the CDS corpus shows a distribution of actor voice and undergoer voice (or passive) which is the reverse from the distribution found in corpora of Tagalog CDS (Garcia et al. 2021; Garcia & Kidd 2022). In line with the observation made in Section 2.4, in the CDS corpus of Jakarta Indonesian, constructions with an actor voice mapping are more frequent than undergoer voice or passive constructions. While 79% of the constructions come in the actor voice (be it unaffixed or prefixed by *N-*), only 21% are either undergoer voice constructions or proper passives (Hidajat 2010: 123–132). There is thus a clear difference between Tagalog-speaking children and Jakarta Indonesian-speaking children with respect to the frequency of each voice in the input that they receive. However, in terms of voice morphology only (i.e. excluding unaffixed predicates), the picture is different. Gil (2006) showed that the voice marker *di-* is more frequent than its counterpart *N-* in CDS. Thus, among affixed forms, undergoer voice predicates are more frequent than actor voice predicates.

In Hidajat's (2010) analysis, undergoer voice constructions and passive sentences (899 in total) are treated together, as they all bear the same verbal marking (the prefix *di-*). Here, however, we separate undergoer voices (with an actor core argument) from passives (with an oblique or omitted actor) and calculate the numbers of each type by means of the data provided in Hidajat (2010: 136) which are split in the different word order patterns. The great majority of clauses occur without any arguments (45%), which does not allow to classify the clauses as either transitive undergoer voice constructions or intransitive passives. The second largest group of *di*-marked predicates (36%) occur with an overt patient argument and no agent. Given that the omission of the agent argument is char-

acteristic of passive constructions and that there seems to be a constraint on zero anaphora in agents of undergoer voice constructions (see Himmelmann 1999 on Tagalog), one could assume that these clauses are passive constructions. Finally, among the clauses in which the agent is overtly realised, 10% are undergoer voice constructions (i.e. with unmarked agents) and 9% are passive constructions (with the agent in a prepositional phrase).

The comparison of Tagalog and Jakarta Indonesian depends on the data considered. In Gil's (2006) study only affixed verb forms are considered and thus the prevalence of *di*-marked forms over *N*-prefixed forms coincides with the Tagalog pattern, in which UV is more frequent than AV. However, if unaffixed predicates are also taken into account, as in Hidajat (2010), the number of sentences in which the subject is an agent argument is much higher than the one in which it is a non-agent argument. This data set therefore shows a distribution opposite to that of Tagalog. A further difference with Tagalog also pertains to the unaffixed forms: unaffixed clauses in Jakarta Indonesian have an actor voice mapping, whereas in Tagalog unaffixed utterances have an undergoer voice mapping. Finally, the Jakarta Indonesian input does not only contain more actor voice but also more passive(-like) constructions than undergoer voice constructions. It might be wondered whether the Jakarta Indonesian input resembles that of a speaker of a language with an active-passive alternation. Nevertheless, as Hidajat (2010: 295) notes, the number of passives/undergoer voice constructions is still high with respect to that of actor voice predicates. While the ratio of passive constructions per hour of speech is around 7–8 in the case of Inuktitut-speaking children (Allen & Crago 1996), 5 or 6 in the case of Sesotho speaking children (Demuth 1989, 1990) and 3 in the case of English-speaking children (Hidajat 2010: 295), the two children in the CDS corpus of Jakarta Indonesian heard, respectively, 12 and 27 passives/undergoer voices per hour. In order to have entirely comparable data (i.e. the ratio of passives per hour only), it would be necessary to extract the undergoer voice constructions from these counts, which is not done in Hidajat (2010). Nevertheless, considering the numbers above, in which it becomes clear that undergoer voice constructions are much less frequent than passives, one can infer that the ratio of passives per hour in Jakarta Indonesian CDS is still high.

3.2.3 *Predictions on Totoli CDS*

Given that the corpora of CDS in both Tagalog and Jakarta Indonesian seem to reflect the voice distribution found in the corpora of adult language, we expect a similar scenario for Totoli. In a corpus of Totoli adult language (Riesberg et al. 2021), clauses in the undergoer voice are more frequent than in the actor voice (70% vs 30%). We thus predict that child-directed speech in Totoli will also be biased towards the undergoer voice. A further reason to expect a high proportion

of clauses with undergoer voice morphology in the input is the fact that imperatives (a typical feature of CDS) are generally expressed with an undergoer voice predicate in Totoli.

Our preliminary observations on the corpus seem to confirm the expected tendencies. For this, we selected two of the transcribed recordings: a 40-minute recording of one of the youngest children (2;0) and a 26-minute recording of one of the oldest (3;6). We annotated the voice-marked predicates produced by the adult (in both cases, the mother of the child) and counted the instances of each voice without excluding repetitions. In both recordings, the distribution of actor voice and undergoer voice very much resembles that of the adult corpus. The input received by the two-year-old child shows a distribution of 71% of undergoer voice forms ($N=114$) and 29% of actor voice forms ($N=46$) while the one received by the four-year-old contains 71% of undergoer voice ($N=48$) and 29% of actor voice clauses ($N=20$). In both recordings, the numbers include both imperative and non-imperative forms.

3.3 Production of voice-marked predicates

Children acquiring a symmetrical voice language need to learn at least two basic transitive constructions which, in most cases, are differentiated only by means of verbal morphology. In terms of production, the questions that arise are, first, when do children start to produce voice-marked predicates? And, second, do children start using the voice morphology of one of the voices before the other? The data that we review in this section shows that the symmetrical voice system is also "symmetrical" in this regard: children seem to start using actor voice and undergoer voice morphology simultaneously. This in turn also indicates an early acquisition of undergoer voice constructions (contrary to what is reported e.g. for passive constructions in English).

3.3.1 *Production in Tagalog*

An analysis of longitudinal spontaneous speech samples from six children showed the emergence of use of actor voice and undergoer voice marking at the same time — at two years of age (Marzan 2013). However, most of children's early verbal productions are not inflected for voice, suggesting that children are still in the process of acquiring voice marking (Garcia & Kidd 2022; Marzan 2013; Marzan et al. 2014). Cross-sectional speech samples from 20 two- to four-year-olds (Garcia & Kidd 2022) showed that production of voice-marked verbs starts becoming more frequent than uninflected verbs around the age of three.

Once children produce mostly voice-marked verbs, they show more frequent use of undergoer voices (61%) than actor voice (39%) (Garcia & Kidd 2022). A patient voice preference was also observed in an experiment. Even in the condition where adult participants preferred to use the actor voice (70%), i.e. to describe action pictures with animate, definite agents and inanimate, indefinite patients (e.g. with verbs like *cut*), children used the actor voice only 45% of the time (Tanaka 2016).

In terms of accuracy in marking the verbs, spontaneous speech samples show that when children inflect verbs with voice, their voice-marking errors are minimal (Barrios & Garcia 2023; Marzan et al. 2014), e.g. adding a locative voice marker (*k**in**abit**an*** for *k**in**abit*) or omitting it (*h**in**awak* for *h**in**awak**an***), using the wrong actor voice marker (***nag**-uubo* for ***um**uubo*). In Garcia and Kidd's corpus, the errors in actor voice (8.1%) and undergoer voices (7.8%) are comparable in proportion. However, more controlled tasks have shown a patient voice advantage in voice-marking. Galang (1982) provided children with an *ang*-phrase prompt (actor, patient) which they had to complete with a voice-marked verb and object to describe a given action picture, e.g. *Ang bata ay…* "The child is…". All groups of children (three-, five-, seven-, and eight-year-olds) showed higher accuracy in producing the patient compared to the actor voice. Note, however, that using the same task, Segalowitz and Galang (1978) found no difference between the actor and patient voice, but five-year-olds performed better than three-year-olds.

A patient voice advantage has also been found in noun marking. Garcia et al. (2018) and Garcia and Kidd (2020) provided three-, five-, and seven-year-olds a voice-marked verb prompt which they should use to describe a picture. In both studies, when given a verb in the actor voice, children made more errors in marking the nouns compared to when they were given a patient-voice verb. The younger children usually dropped one of the noun markers, but the older children switched the *ang* and the *ng* markers. These noun marker reversals indicate that children might have been using the patient voice noun mapping by default, which resulted in reversals in the actor voice.

These findings show that at the age of three years, children start inflecting verbs with voice, and they usually do not make errors in their productions. However, given more challenging tasks, such as in experiments where their linguistic skills are tested, children show higher accuracy in using the patient voice compared to the actor voice. This patient voice advantage is also observed in the frequency of their voice productions. In line with the input that they receive, children also produce more verbs in the undergoer than the actor voice in spontaneous speech.

3.3.2 *Production in Jakarta Indonesian*

Recall that the data on CDS from Jakarta Indonesian reveal a high frequency of unmarked predicates with an actor voice alignment (Hidajat 2010) although, when only affixed forms are observed, predicates prefixed by *di-* (the undergoer voice or passive marker) are more frequent than the ones marked by the actor voice predicate *N-* (Gil 2006). Despite this being different to the Tagalog case, where unaffixed verbs follow the patient voice noun marking pattern and where affixed undergoer voice constructions are more frequent than actor voice forms, the studies on the acquisition of Jakarta Indonesian show that children also acquire undergoer voice and passive predicates from an early age (Gil 2006; Hidajat 2010).

Gil (2006) analyses the production of voice morphology in a longitudinal corpus of eight children between 1;6 and 7;9 years of age. The study focuses on the two voice marking prefixes *di-* and *N-* and shows that children start to use both prefixes productively and in a correct manner from an early age, sometimes even under 2;0. As reported for Tagalog above, Gil also reports that actor and undergoer voice marking emerges at the same time.

Hidajat (2010) also analysed the production by two out of the eight children of the corpus used by Gil (2006) but also included unaffixed predicates. The results showed that the earliest and most frequent productions by the two children are clauses with unaffixed verbs and actor voice configuration. However, among voice-marked forms, *di*-predicates are produced more frequently than *N*-predicates. From a total of 1.718 predicates by one of the children, 1.155 are in the unaffixed form (67%), 427 predicates (including imperatives) are marked by *di-* (25%), and 136 are marked by the actor voice prefix *N-* (8%). From the 3074 sentences of the other child, 2005 are unaffixed (65%), 828 predicates (including imperatives) are prefixed by *di-* (27%), and 241 are prefixed by *N-* (8%). Among the predicates marked by *di-*, constructions without an agent are the most frequent productions, followed by passives with an overt agent in a prepositional phrase. Undergoer voice constructions (with an overt unmarked agent) are the least frequent (Hidajat 2010: 227).

These proportions reveal both differences and similarities with the Tagalog data. One of the differences pertains again to unaffixed forms: while in Tagalog (although not used in adult-directed speech) they have undergoer voice alignment, in Hidajat's (2010) Indonesian data, unaffixed predicates have an actor voice alignment and show a high frequency. The Tagalog and the Jakarta Indonesian data coincide if only affixed forms are considered: children produce more predicates in which the undergoer is the subject (UV/passives) than predicates with agent subjects (AV). A further difference, however, is that in the Jakarta Indonesian data reported by Hidajat, *di*-marked forms also include passives, which are

not possible in Tagalog. Jakarta Indonesian children thus show an early production of 4 different structures: two with actor voice alignment (unmarked or prefixed by *N-*), a passive one and a transitive undergoer voice form.

3.3.3 *Predictions on production in Totoli*

Given the data on the use of voice morphology in Tagalog and Jakarta Indonesian, we also expect that both actor voice and undergoer voice morphology emerge at the same time. Considering the findings of Tagalog, a phase in which children produce unaffixed verbs is also plausible.

An observation of the two recordings commented on above also suggests a simultaneous emergence of actor voice and undergoer voice morphology. In both recordings, non-verbal forms such as existential or equational clauses seem to be very frequent, but verbal predicates are also produced. We observe that not only the child at age 3;6 but also the two-year-old child correctly used actor voice and undergoer voice morphology in verbal predicates. This is illustrated in Examples (7)–(10). Although examples from the two-year-old still differ from the targeted word, the relevant segments (*-i* in the UV and the nasal in the AV) are present. In the 40-minute recording, the two-year-old child produced the same amount of actor-voice-marked and undergoer-voice-marked predicates (15 of each). By contrast, the child at age 3;6 produced more predicates marked for undergoer voice than for actor voice (21 vs. 4).

(7) UV clause (2;0)
CHILD: *gia*
PRX
'this' (showing a broken toy)
CHILD: *tei ia*
meant: *koot-i(=ai) ia?*
cut-UV.APPL(=VEN) PRX
'it was cut'
MOT.: *koot-i=ai ia?*
cut-UV.APPL=VEN PRX
'was it cut?'
CHILD: *iye*
'yes'

(8) AV clause (2;0)
MOT.: *ngina bau ana?*
where? fish MED
'where is that fish?'

CHILD: *ina nam bini*
meant: *ina mangaan bini*
ina moN-kaan bini
mother AV-eat rice
'mum eats rice'

(9) UV clause (3;6)
CHILD: *kombo nikootna ia?*
kombula ni-koot=na ia
why? UV.RLS-cut=3SG.GEN PRX
'why did he cut this?'

(10) AV clause (3;6)
CHILD: *mogutupo kukuku*
mo-gutu=po kuku=ku
AV-do=ICPL foot=1SG.GEN
'(I) do my feet first' (the child wants to paint her nails)

With regard to unaffixed predicates, the data is thus far inconclusive. In the productions of the two-year-old child, several verbs seem unaffixed, especially the ones which should bear a prefix, and less so if they should take a suffix. Nevertheless, we observe that this child often truncates words (mostly word-initially), which could explain the absence of prefixes. The child at age 3;6 does not seem to produce unaffixed predicates.

3.4 Use of morphosyntactic markers for thematic role assignment

Speakers of symmetrical voice languages also need to learn to use morphology to assign thematic roles to the arguments of a clause. Unlike in an asymmetrical alternation, children face the additional challenge that they cannot assign the agent role to the subject argument by default. In this regard, there are two different scenarios. In Philippine-type languages, children have to map voice morphology and phrase marking clitics to assign thematic roles. By contrast, in Indonesian-type languages and North and Central Sulawesi languages, nouns are unmarked, so children have to rely on voice morphology and word order.

3.4.1 *Thematic role assignment in Tagalog*

In line with the high frequency of undergoer voice and of non-actors as subjects in Tagalog CDS, psycholinguistic experiments have shown a patient voice advantage in children's use of the voice-marking on the verb and the noun markers for thematic role assignment. Using a picture selection task, Segalowitz and Galang

(1978) found that three-, five-, and seven-year-old children correctly interpreted patient voice (verb-actor-patient, VAP) sentences. However, for actor-voice (verb–patient–actor, VPA) sentences, all age groups incorrectly chose the picture in which the noun mentioned first was the doer of the action. A follow-up experiment with only the five- and seven-year-olds using verb-medial sentences (actor-verb-patient AVP in the actor voice and patient-verb-actor PVA in the patient voice) showed high accuracy for both voices, even though the patient voice sentences were patient-initial. Both experiments together show that children used the morphosyntactic markers for assigning thematic roles in the patient voice, but that they relied on a first-noun-as-actor strategy in the actor voice. The same patient voice advantage was found by Galang (1982) who asked children to identify the referent of the *ang*-phrase, e.g. *Ituro mo ang k**um**akain/k**in**akain*, 'Point to that which is eating/being eaten'. Even the three-year-olds were accurate in following instructions with the patient voice *-in-* marker. Three- and five-year-old children were at chance for the actor voice.

More recently, Garcia et al. (2019) found similar results using voice-marked verb-initial sentences where the order of agent and patient was modulated (AV VAP, AV VPA, PV VAP, PV VPA). Through a self-paced listening and picture verification task, they found that five- and seven-year-olds were equally accurate in interpreting VAP and VPA sentences in the patient voice, but in the actor voice they performed better in VAP compared to VPA sentences. They also had longer listening times given sentence-picture mismatches (i.e. showing thematic role reversal) compared to sentence-picture matches, only in the patient voice, indicating that they detected mismatches in noun marking only in this voice. In a picture-selection task with eye-tracking (Garcia et al. 2020), children showed above chance accuracy for both orders in the patient voice and for the actor voice VAP condition. However, for the actor voice VPA condition, seven-year-olds scored at chance level, and five-year-olds scored below chance; indicating a first-noun-as-actor strategy in the actor voice. A patient voice advantage was also observed in using the morphosyntactic markers for predicting thematic roles. Garcia et al. (2021) found that, like adults, seven- and nine-year-olds used the voice-marking on the verb and the first noun marker to predict the first noun, but only in the patient voice and not in the actor voice.

Overall, the findings show that when there is no other available cue for sentence interpretation (e.g. animacy, real-world probability), children do not show adult-like use of the voice-marking on the verb and the noun markers for assigning thematic roles even at the age of seven years. The studies also point to a patient voice advantage, in that children show more adult-like comprehension when given sentences with a verb in the patient voice compared to the actor voice.

3.4.2 *Thematic role assignment in Jakarta Indonesian*

Despite a higher frequency of unaffixed forms with an actor voice configuration both in the input and in the production of Jakarta Indonesian children, the work by Hidajat (2010) suggests that children do not show more difficulties in the comprehension of undergoer voice/passive forms than with unaffixed predicates with actor voice alignments. The experimental data by Hidajat (2010) tested the comprehension of unaffixed actor voice clauses and undergoer voice clauses by means of picture selection tasks. Test items with *N*-marked actor voice forms were not included, which makes the data less comparable to the Tagalog one. In two experiments conducted on children between 2;9 and 4;0 years, Hidajat (2010) tested the comprehension of familiar verbs (experiment 1) and novel verbs (experiment 2). All test items had an SVO order (i.e. AVP in the unaffixed verbs and PVA in the undergoer voice) and arguments were realized as full noun phrases. In both experiments, children scored above chance in all conditions. These results indicate that children were able to use voice-marking on the verb to correctly assign the thematic roles in the undergoer voice (PVA order), instead of relying on a first-noun-as-actor strategy. The results also showed that performance on items containing unaffixed verbs in actor voice was numerically higher than performance on items containing verbs in the undergoer voice, but the difference did not reach conventional standards of statistical significance (Hidajat 2010: 319–352). Two further experiments by Hidajat (2010) tested unaffixed actor voice clauses and undergoer voice/passive constructions in which one of the two arguments was omitted. In both, children showed good comprehension of both voices. In fact, scores were slightly higher for the [subject+*di*-verb] condition than for [*di*-verb+object], [subject+unaffixed verb] and [unaffixed verb+object].

The above chance accuracy in all conditions indicates that children already have good command of the two voices (even when one argument is omitted) before the age of 3;0 and thus can use verbal morphology for thematic role assignment (Hidajat 2010). However, Jakarta Indonesian nouns are unmarked, unlike in Tagalog. Children thus do not use nominal morphology for thematic role assignment. Rather they have to map verbal morphology and word order. Nevertheless, the Jakarta Indonesian data available does not allow us to make any observation with regard to the role of word order in thematic role assignment independent of voice, although the successful early use and comprehension of all voice types suggests that children do not solely rely on word order cues to assign thematic roles, as has been observed in European languages like English and German (e.g. Dittmar et al. 2008).

3.4.3 *Predictions on thematic role assignment in Totoli*

As in Jakarta Indonesian, Totoli nouns are not marked for case. Children thus have to use voice morphology and word order for thematic role assignment. The subject argument can either precede the verb (SVO) or follow the verb+object phrase (i.e. VOS). This means that in SVO sentences, upon hearing the first noun, its thematic role is still ambiguous. It is not until the verb is heard that children can unambiguously assign thematic role. The main question here will be whether children appropriately use the voice-marking on the verb to correctly assign the thematic roles in the undergoer voice (PVA order), instead of relying on a first-noun-as-actor strategy. The Indonesian data suggests that children do not rely on a first-noun-as-actor strategy in the undergoer voice and this could be the case in Totoli too. Nevertheless, there is a difference between Jakarta Indonesian and Totoli. The CDS data in Hidajat (2010) indicates that PVA orders in passives and undergoer voice predicates are frequent. In contrast, while a PVA order is possible in Totoli undergoer voices, adult data show that a VAP order is much more frequent (Riesberg et al. under review). Additionally, in the actor voice, AVP is much more frequent than VPA. Thus, pre-verbal nouns in Totoli are generally actors and not patients, so it cannot be ruled out that Totoli children rely on a first-noun-as-actor strategy. This shows that the corpus data obtained from language documentation unfortunately cannot answer all relevant questions. This particular topic would need to be covered by means of experimental studies.

3.5 Word order in children's productions of symmetrical voice structures

While the three types of symmetrical voice languages allow for different word order patterns, there is a general preference for agent-first orders among western Austronesian languages (Riesberg et al. 2019). This, together with the claim that agent-first orders are easier to process (Bornkessel & Schlesewsky 2006; Hawkins 2014), raises the question of whether an agent-first bias can also be observed in the acquisition of a symmetrical voice language. In terms of comprehension/thematic role assignment, the last section showed that children do not consistently apply a first-noun-as-actor strategy. However, this does not exclude an agent-first bias in their production, as observed in adults. The results reviewed below suggest that this preference is found to different extents in Tagalog (agent-first bias in both voices) and Jakarta Indonesian (agent-first bias only in actor voices). As for Totoli, data is still scarce but a scenario closer to the Tagalog one seems more plausible.

3.5.1 *Word order in Tagalog*

An analysis of spontaneous speech samples from two- to four-year-olds (Garcia & Kidd 2022) shows that utterances with two arguments are typically actor-first, regardless of voice (AV: 86%, PV: 75%, uninflected: 89%). This first argument (actor) is also typically a pronoun (AV: 77%, PV: 90%, uninflected: 92%). Utterances with at least one pronoun have a more rigid word order, in that the pronoun must occur immediately after the predicate (Billings 2005). When one argument is dropped, the remaining argument is usually still the actor for the actor voice and uninflected verbs, but not for the patient voice (AV: 88%, PV: 48%, uninflected: 71%). Note though that across all voices, the only argument is typically a pronoun (AV: 76%, PV: 74%, uninflected: 77%).

In a sentence completion task, Garcia et al. (2018) asked five- and seven-year-olds to describe a picture of an action between two animals, by completing a sentence that started with a given voice-marked verb prompt. The children were asked to use full noun phrases. Results showed that for both actor and patient voice, children produced mostly actor-initial sentences. This result differs from the adult productions, which were mostly actor-initial for the patient voice, but not for the actor voice, which showed almost an equal distribution of actor-initial and patient-initial sentences. The latter results are in line with the proposed principles of ordering full noun phrases in Tagalog, namely actor-before-patient and subject-last (aside from light-before-heavy; Kroeger 1993). In the patient voice, both of these principles are satisfied by an actor-initial sentence. However, in the actor voice, these principles are in opposition, leading to no preference of a particular order. The children's productions indicate that children prioritise the actor-first over the subject-last principle.

3.5.2 *Word order in Indonesian*

The corpus analysis of the spontaneous production by two Jakarta Indonesian children between 2;0 and 3;0 in Hidajat (2010) reveals an agent-first preference in clauses with an actor voice configuration. In general, however, clauses with overt agent and patient arguments are not very frequent. In the productions of unaffixed actor voice clauses, orders in which the agent precedes the patient (60 out of 1.155 different orders by one child and 236 out of 2.005 by the other) are more frequent than orders in which the agent follows the patient (12 by one child, 86 by the other). Apart from this set of unaffixed predicates, all other form types occur too infrequently in spontaneous data to identify robust tendencies. Actor voices marked with *N-* are overall infrequent and even more with two overt arguments. One child only produced three of such clauses, two with an agent-before-patient word order and another one with the patient preceding the agent. The

second child, however, shows a clear preference of agent-first clauses (28) over patient-before-agent clauses (only 1 instance). As for passive or undergoer voice predicates, numbers are also small, but the ones available, unlike in Tagalog, do not show a preference for agent-first orders. One of the two children produced 13 clauses with a patient-first argument and only 2 clauses with an agent-first argument (the other child is not considered, as it only produced one sentence with overt agent and patient).

3.5.3 *Predictions on word order in Totoli*

In Totoli, SVO and VOS word orders are possible in all voices. However, in adult corpora, in 80% of the cases the word order pattern follows an agent-first bias: i.e. SVO in the actor voice (= AVP) and VOS in the undergoer voice (= VAP) (Riesberg et al. under review). On this basis, one could expect that both CDS and speech produced by children also show an agent first preference. In the two recordings analysed, the input received by children rarely includes a clause with two overt arguments, as either one or two arguments tend to be realized as zero. The number of transitive clauses with overt agent and patient is thus very small (18 in one recording and 12 in the other). Nevertheless, in both cases the number of clauses in which the agent precedes the patient is higher than the number of clauses in which the patient precedes the agent. The child at age 2;0 heard 12 clauses with an agent-first order and 6 clauses with the patient before the agent. The child at age 3;6 heard 10 clauses with the agent preceding the patient and 2 clauses with the agent following the patient. The two recordings do not provide enough data for a first observation on children's production: the two-year-old child did not produce any clause with overt agent and patient and the older child produced only six, five of them with an agent-first order. Given the low occurrence of transitive clauses with two overt arguments in spontaneous speech (both by the caregiver and the child), an investigation of agent-first preference in Totoli would profit from an experimental approach.

4. Conclusions

This paper compared corpus data and experimental data with regard to the acquisition of the symmetrical voice system of Tagalog (a Philippine-type language), Jakarta Indonesian (an Indonesian-type language) and Totoli (a North and Central-Sulawesi type language). The comparison enabled us to investigate whether the typological division of symmetrical voice languages is also reflected in the acquisition of the symmetrical voice system. One of the main differences observed concerns the frequency of each voice in CDS. With the data at hand, it seems that the

input heard by children in a Philippine-type language and potentially also in a language from North and Central Sulawesi is biased towards the undergoer voice. By contrast, in an Indonesian-type language such as Jakarta Indonesian, clauses with an actor voice configuration are more frequent than passives or undergoer voice clauses when both affixed and unaffixed verbs are counted. If only prefixed verbs are considered, passive/undergoer voice forms are more frequent than actor voice predicates. It remains to be investigated how frequencies appear in an Indonesian-type language that does not allow for unmarked actor voice predicates.

Similar frequencies are observed in children's production. Tagalog children produce more correct undergoer voice than actor voice predicates. By contrast, Jakarta Indonesian children produce more unmarked actor voice forms than undergoer voice or passive forms, which, in turn, are more frequently produced than *N*-marked actor voice clauses. Totoli data is still to be analysed but it is predicted that production will be biased towards the undergoer voice. Despite these differing frequencies, spontaneous data of all three languages do not show a difference in the age of acquisition of each construction. In fact, the corpus data suggests that children start to produce actor voice and undergoer voice morphology at the same time. This indicates that in all three types, both the construction in which the agent is the subject argument and the construction in which it is not are acquired from an early age, although it must be acknowledged that the early use of different voice constructions does not necessarily equate to full mastery, as shown by experimental work (e.g. Garcia et al. 2019; Garcia & Kidd 2020). This raises the possibility that the early use of different voice constructions may be tied to individual verbs that prefer one frame over another, following lexicalist approaches to the early acquisition of argument structure (e.g. Pine et al. 1998; Tomasello 2000).

Further differences are found with regard to thematic role assignment and the agent-first preference. The Tagalog findings show that children used the morphosyntactic markers for thematic role assignment in the patient voice, but that they relied on a first-noun-as-actor strategy in the actor voice. This agent-first preference is also observed in the production of sentences in both voices. Jakarta Indonesian children also correctly used voice morphology to assign thematic roles in the undergoer voice instead of a first-noun-as-actor strategy. In the actor voice, all items tested have an agent-first order and therefore no claims on word order can be made independently of voice morphology. In terms of production, an agent-first preference is found in the actor voice but not in the undergoer voice. As for Totoli, an agent-first preference is expected in both voices (given that this is found in adult data) so that a first-noun-as-actor strategy cannot be ruled out.

Beyond the theoretical implications, the current chapter also showcases the importance of well-annotated and broad language corpora for theory building. While Himmelmann (1998, 2006) already described the goal of language

documentation to be as all-encompassing as possible, this has only rarely been extended to include child language and child-directed language. A growing number of language corpora of western Austronesian languages are available, yet data on language acquisition is rare. In Himmelmann's (1998, 2006) definition of the goal of language documentation, the inclusion of such data as an important part of human language is imperative. In the context of the Sketch Acquisition Manual (Hellwig et al. 2023), Bracks and Bardají (2023) have highlighted this point and discussed the process and the hurdles of the inclusion of such data, exemplifying with their own experience in the documentation of Totoli. While the Totoli data was not collected with the specific aim of analysing the acquisition of symmetrical voice, the present chapter demonstrates how the inclusion of such data types in documentation projects enables research on topics that were not targeted at the moment of data collection.

Acknowledgements

We would like to thank the Totoli and Tagalog speakers who participated in the recordings and experiments mentioned in this paper. Many thanks also to Lanny Hidajat for sharing her doctoral dissertation with us. We also gratefully acknowledge the financial support from the Deutsche Forschungsgemeinschaft (CRC 1252 "Prominence in Language") in the data collection.

More than anyone we would like to thank Nikolaus Himmelmann for being present in multiple ways, for laying the ground at so many levels, for enriching our work with his reflections and knowledge, for helping us grow and, most especially, for being such a kind supervisor, colleague and friend.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following abbreviations are used:

AV actor voice
CV conveyance voice
HON honorific
i inclusive
ICPL incompletive
LK linker
LV locative voice
MED medial
PN personal name
PRF perfective
PV patient voice
RDP reduplication
RLS realis
SF stem former
SPEC specifier
UV undergoer voice
VEN venitive.

References

Adricula, Norielle. 2023. "Examining Voice Choice in Tagalog: A Corpus of Web-Based Tagalog." *Proceedings of the Linguistic Society of America* 8 (1): 5533.

Allen, Shanley E. M., and Martha B. Crago. 1996. "Early Passive Acquisition in Inuktitut." *Journal of Child Language* 23: 129–155.

Arka, I Wayan. 2003. "Voice Systems in the Austronesian Languages of Nusantara: Typology, Symmetricality and Undergoer Orientation." *Linguistik Indonesia* 21: 113–139.

Arka, I Wayan, and Jane Simpson. 2008. "Objective Voice and Control into Subject Clauses in Balinese." In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter K. Austin, and Simon Musgrave, 90–127. Stanford: CSLI Publications. http://hdl.handle.net/1885/26472

Bardají, Maria. 2024. *Nominalization in Western Austronesian: Not-at-Issueness in Totoli and Beyond*. Berlin, Boston: De Gruyter Mouton.

Barrios, Aireen, and Rowena Garcia. 2023. "Filipino Children's Acquisition of Nominal and Verbal Markers in L1 and L2 Tagalog." *Languages* 8 (3): 188.

Bautista, Ma. Lourdes S. 1983. "Word Order in Tagalog Child Language." *Philippine Journal of Linguistics* 14: 25–44.

Billings, Loren A. 2005. "Ordering Clitics and Postverbal R-Expressions in Tagalog: A Unified Analysis?" In *Verb First: On the syntax of verb-initial languages*, ed. by Andrew Carnie, Heidi Harley, and Sheila Dooley, 303–339. Amsterdam: John Benjamins.

Bornkessel, Ina, and Matthias Schlesewsky. 2006. "The Extended Argument Dependency Model: A Neurocognitive Approach to Sentence Comprehension Across Languages." *Psychological Review* 113 (4): 787–821.

Bracks, Christoph A., and Maria Bardají. 2023. "Broadening Language Documentation with Child Language Data: First-Hand Experience from Tolitoli." In *The Acquisition Sketch Project*, ed. by Birgit Hellwig, Shanley E. M. Allen, Lucinda Davidson, Rebecca Defina, Barbara F. Kelly, and Evan Kidd, 87–108. Honolulu: University of Hawai'i Press. https://hdl.handle.net/10125/74721

Bracks, Christoph, and Nikolaus P. Himmelmann. In preparation. "How Long Does It Take for a Language to Disappear? Rapid and Radical Language Shift in the Village of Gio (Sulawesi, Indonesia)."

Bracks, Christoph A., Datra Hasan, Maria Bardají i Farré, Sumitro Pogi, and Nikolaus P. Himmelmann. 2023. *Totoli Documentation Corpus 2*. Data Center for the Humanities.

Brickell, Timothy C. 2014. A Grammatical Description of the Tondano (Toundano) Language. Doctoral Dissertation, La Trobe University.

Chen, Victoria, and Bradley McDonnell. 2019. "Western Austronesian Voice." *Annual Review of Linguistics* 5 (1): 173–95.

Cumming, Susanna. 1991. *Functional Change: The Case of Malay Constituent Order*. Berlin, New York: De Gruyter Mouton.

Davies, William D. 2010. *A Grammar of Madurese*. Berlin, New York: De Gruyter Mouton.

Defina, Rebecca, Shanley E. M. Allen, Lucinda Davidson, Birgit Hellwig, Barbara F. Kelly, and Evan Kidd. 2023. "Sketch Acquisition Manual (SAM), Part II: The Acquisition Sketch." In *The Acquisition Sketch Project*, ed. by Birgit Hellwig, Shanley E. M. Allen, Lucinda Davidson, Rebecca Defina, Barbara F. Kelly, and Evan Kidd, 39–86. Honolulu: University of Hawai'i Press.

Demuth, Katherine. 1989. "Maturation and the Acquisition of the Sesotho Passive." *Language* 65, 56–80.

Demuth, Katherine. 1990. "Subject, Topic and Sesotho Passive." *Journal of Child Language* 17 (1): 67–84.

Dittmar, Miriam, Kirsten Abbot-Smith, Elena Lieven, and Michael Tomasello. 2008. "German Children's Comprehension of Word Order and Case Marking in Causative Sentences." *Child Development* 79: 1152–1167.

Fatinah, Siti, Ryen Maerina, Heksa Biopsi Puji Hastuti, Deni Karsana, Nuraidar Agus, Dwiani Septiana, and Andi Indah Yulianti. 2024. "Totoli Language: Could It Be Extinct?" *Cogent Arts and Humanities* 12 (1): 1–28.

Galang, Rosita G. 1982. "Acquisition of Tagalog Verb Morphology: Linguistic and Cognitive Factors." *Philippine Journal of Linguistics* 13: 1–15.

Garcia, Rowena. 2019. "Child Language Corpus" in *Tagalog Corpus*. The Language Archive. https://hdl.handle.net/1839/43d11ac1-c666-4249-92cb-afd88387caaf. Accessed May 4, 2025.

Garcia, Rowena, and Evan Kidd. 2020. "The Acquisition of the Tagalog Symmetrical Voice System: Evidence from Structural Priming." *Language Learning and Development* 16 (4): 399–425.

Garcia, Rowena, and Evan Kidd. 2022. "Acquiring Verb-Argument Structure in Tagalog: A Multivariate Corpus Analysis of Caregiver and Child Speech." *Linguistics* 60 (6): 1855–1906.

Garcia, Rowena, Jens Roeser, and Barbara Höhle. 2019. "Thematic Role Assignment in the L1 Acquisition of Tagalog: Use of Word Order and Morphosyntactic Markers." *Language Acquisition* 26 (3): 235–61.

Garcia, Rowena, Jens Roeser, and Barbara Höhle. 2020. "Children's Online Use of Word Order and Morphosyntactic Markers in Tagalog Thematic Role Assignment: An Eye-Tracking Study." *Journal of Child Language* 47 (3): 533–55.

Garcia, Rowena, Gabriela Garrido Rodriguez, and Evan Kidd. 2021. "Developmental Effects in the Online Use of Morphosyntactic Cues in Sentence Processing: Evidence from Tagalog." *Cognition* 216 (November): 104859.

Garcia, Rowena, Jeruen E. Dery, Jens Roeser, and Barbara Höhle. 2018. "Word Order Preferences of Tagalog-Speaking Adults and Children." *First Language* 38 (6): 617–640.

Garcia, Rowena, Michael C. Valdez, Annika McDermott-Hinman, Scott AnderBois, and Roman Feiman. In preparation. *A Longitudinal Corpus of Tagalog Child-Directed and Children's Speech*.

Gil, David, and Uri Tadmor. 2007. *The MPI-EVA Jakarta Child Language Database*. Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and Center for Language and Culture Studies, Atma Jaya Catholic University.

Hawkins, John A. 2014. *Cross-Linguistic Variation and Efficiency*. Oxford: Oxford University Press.

Hellwig, Birgit, Shanley E. M. Allen, Lucinda Davidson, Rebecca Defina, Barbara F. Kelly, and Evan Kidd (eds). 2023. *The Acquisition Sketch Project*. Honolulu: University of Hawai'i Press. https://hdl.handle.net/10125/104899

Hidajat, Lanny. 2010. The Acquisition of Verb Argument Structure in Basilectal Jakarta Indonesian. Doctoral Dissertation, University of Delaware.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 1999. "The Lack of Zero Anaphora and Incipient Person Marking in Tagalog". *Oceanic Linguistics* 38 (2): 231–269.

Himmelmann, Nikolaus P. 2002a. "Voice in Western Austronesian: An Update." In *The History and Typology of Western Austronesian Voice Systems*, ed. by Fay Wouk, and Malcolm Ross, 7–17. Canberra: Pacific Linguistics.

Himmelmann, Nikolaus P. 2002b. "Voice in Two Northern Sulawesi Languages." In *The History and Typology of Western Austronesian Voice Systems*, ed. by Fay Wouk, and Malcolm Ross, 123–142. Canberra: Pacific Linguistics.

Himmelmann, Nikolaus P. 2005a. "The Austronesian Languages of Asia and Madagascar: Typological Characteristics." In *The Austronesian Languages of Asia and Madagascar*, ed. by Alexander Adelaar, and Nikolaus P. Himmelmann, 110–181. London: Routledge.

Himmelmann, Nikolaus P. 2005b. "Tagalog." In *The Austronesian Languages of Asia and Madagascar*, ed. by Alexander Adelaar, and Nikolaus P. Himmelmann, 350–376. London: Routledge.

Himmelmann, Nikolaus P. 2006. "Language Documentation: What is It and What is It Good for?", in *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin, New York: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2008. "Lexical Categories and Voice in Tagalog." In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter K. Austin, and Simon Musgrave, 247–293. Stanford: CSLI Publications.

Himmelmann, Nikolaus P., and Sonja Riesberg. 2013. "Symmetrical Voice and Applicative Alternations: Evidence from Totoli." *Oceanic Linguistics* 52 (2): 396–422.

Huber, Eva, Aireen Barrios, Rowena Garcia, Dagmar Jung, Evan Kidd, Balthasar Bickel, and Sabine Stoll. 2025. Child Language Corpus of Batangas Tagalog (Version 1.0) [Data set]. LaRS - Language Repository of Switzerland.

Jap, Bernard. A. J., and Yu-Yin Hsu. 2025. "An ERP Study on Verb Bias and Thematic Role Assignment in Standard Indonesian." *Scientific Reports* 15: 11847.

Kidd, Evan, and Rowena Garcia. 2022. "How Diverse Is Child Language Acquisition Research?" *First Language* 42: 703–735.

Kroeger, Paul R. 1993. *Phrase Structure and Grammatical Relations in Tagalog*. Stanford: CSLI Publications.

Leto, Claudia, Winarno S. Alamudi, Nikolaus P. Himmelmann, Jani Kuhnt-Saptodewo, Sonja Riesberg, and Hasan Basri. 2005–2010. *DoBeS Totoli documentation*. The Language Archive. https://hdl.handle.net/1839/da11addf-bef3-4742-9c00-d85a446f2cdb

Marzan, Jocelyn Christina B. 2013. Spoken Language Patterns of Selected Filipino Toddlers and Preschool Children. Doctoral Dissertation, Quezon City, Philippines: University of the Philippines Diliman.

Marzan, Jocelyn Christina B., Maryelle Ellaine L. Dy, Mary Seanne A. Angeles, Jessica Francesca P. Arnaldo, Ma. Bettina Clare N. Camacho, Thea Beatriz G. Cruz, Ancellie L. Espino, Johannes C. Ibardaloza, Doris Renee V. Leaño, Chef Stan L. Macaraeg, Anna Christen D. L. Magbanua, and Grace Camille S. Tan. 2014. Correctness of Verb Inflections Coding Aspect and Tense Used by Five 14 to 44-Month-Old Filipino Children: An Exploratory Descriptive Longitudinal Study. B.A. Thesis, Manila, Philippines: University of the Philippines Manila.

McDonnell, Bradley. 2016. Symmetrical Voice Constructions in Besemah: A Usage-Based Approach. Doctoral Dissertation, University of California, Santa Barbara.

doi Passmore, Sam, Birgit Hellwig, Rowena Garcia, and Evan Kidd. 2025. "The Scientific and Cultural Cost of Convenience Sampling in the Face of Rising Language Endangerment: Highlighting the Role of Language Acquisition." *Open Mind* 9: 501–514.

Pastika, I Wayan. 1999. Voice Selection in Balinese Discourse. Doctoral Dissertation, Australian National University.

doi Pine, Julian M., Elena V. M. Lieven, and Caroline F. Rowland. 1998. "Comparing Different Models of the Development of the English Verb Category". *Linguistics* 36 (4): 807–830.

doi Pine, Julian, Joanna Kolak, Virve Vihman, Felix Engelmann, Sonia Granlund, Anna Theakston, Elena V. M. Lieven, Judit Fazekas, and Ben Ambridge. 2025. "Why Learners Privilege Word-Order over Case-Marking: A Cross-Linguistic Meta-Analysis, New Data from Estonian, Finnish and Polish, and a Discriminative Learning Model." *Psychological Review*.

doi Pinker, Steven. 1984. *Language Learnability and Language Development*. Cambridge: Harvard University Press.

doi Pinker, Steven. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. The MIT Press.

Quick, Phil. 2005. "Topic Continuity, Voice and Word Order in Pendau." In *The Many Faces of Austronesian Voice Systems: Some New Empirical Studies*, ed. by I Wayan Arka, and Malcolm Ross, 221–242. Canberra: Pacific Linguistics.

doi Riesberg, Sonja. 2014. *Symmetrical Voice and Linking in Western Austronesian Languages*. Berlin, Boston: De Gruyter Mouton.

doi Riesberg, Sonja, Kurt Malcher, and Nikolaus P. Himmelmann. 2019. "How Universal Is Agent-First? Evidence from Symmetrical Voice Languages." *Language* 95 (3): 523–561.

doi Riesberg, Sonja, Maria Bardají i Farré, Kurt Malcher, and Nikolaus P. Himmelmann. 2021. "Predicting Voice Choice in Symmetrical Voice Languages: All the Things That Do not Work in Totoli." *Studies in Language* 46 (2): 453–516.

Riesberg, Sonja, Mark. T. Ellison, Maria Bardají, and Nikolaus P. Himmelmann. Under review. *When Agents Are not Subjects — What Symmetrical Voice Languages Can Tell Us about Word Order Variation*.

doi Segalowitz, Norman S., and Rosita G. Galang. 1978. "Agent–Patient Word-Order Preference in the Acquisition of Tagalog." *Journal of Child Language* 5 (1): 47–64.

doi Seifart, Frank, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. "Language Documentation Twenty-Five Years on." *Language* 94 (4): e324–e345.

Tanaka, Nozomi. 2016. An Asymmetry in the Acquisition of Tagalog Relative Clauses. Doctoral Dissertation, University of Hawaii at Manoa.

Tomasello, Michael. 2000. "Do Young Children Have Adult Syntactic Competence?". *Cognition* 74 (3): 209–253.

Wolff, John U. 1996. "The Development of the Passive Verb with Pronominal Prefix in Western Austronesian Languages". In *Reconstruction, Classification, Description — Festschrift in Honor of Isidore Dyen*, ed. by Bernd Nothofer, 15–40. Hamburg: Abera-Verlag.

Discourse structure

CHAPTER 23

On syntactic uniformity and the functional effects of verbal DPs

Katharina Haude
Structure et Dynamique des Langues (CNRS UMR8202 — INALCO — IRD)

According to Himmelmann's (2008: 249) Syntactic Uniformity Hypothesis for Tagalog, words from different lexical categories "may all occur in essentially the same basic syntactic positions". This is also true of Movima (isolate, Bolivia). Based on a documentation corpus of spontaneous Movima discourse data, this paper furthermore shows that argument DPs containing a verb are particularly frequent in combination with nonverbal predicates (nouns, quantifiers, demonstratives), i.e. their syntactic distribution is diametrically opposed to that of nominal DPs. It is suggested that in a system with syntactic uniformity, a DP containing a verb functions in a way similar to a headless or light-headed relative clause in a language whose verbs and nouns are more strictly linked to the syntactic categories predicate and argument.

Keywords: lexical categories, omnipredicativity, relative clauses, nominalization, nonverbal predicates, Movima, discourse, focus, existential constructions

1. Introduction

Based on the well-known fact that in Tagalog, all lexical roots and derived words can in principle occur as predicates and in determiner phrases (DPs) alike, Himmelmann formulates a "Syntactic Uniformity Hypothesis" for Tagalog:

> Syntactic Uniformity Hypothesis:
> "[A]ll content words may occur, without further derivation or conversion, in the same kind of phrase-structural positions." (Himmelmann 2008: 267)

The value of this proposal lies in the fact that it makes the occurrence of a content word in a particular syntactic position independent of the lexical category of the

https://doi.org/10.1075/slcs.240.23hau
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

content word. Tagalog nouns can function directly as predicates, without the need of a copula; this is illustrated in (1). And a word occurring in a DP does not have to be a noun: it can just as well be a morphological verb, as in (2), where the action-denoting, voice-marked verb occurs in a DP (in square brackets) marked by the definite article *ang*.

(1) Tagalog (Himmelmann 1991: 44)
mang-ga-gamot siyá
IRR.ACT-RED2-medicine 3.SG
'He is a doctor.'

(2) Tagalog (Himmelmann 2008: 287)
yung anák [ang sumigáw]
iyón.LK anák ang um-sigáw
DIST.LK child SPEC AV-shout
'The one who shouted/the shouter is (was) that child.'

This phenomenon has been observed in other, genealogically and areally unrelated languages, in particular Salish languages (Kinkade 1983; Jelinek & Demers 1994), Classical Nahuatl (Launey 1994; 2004), and Movima (Haude 2009).

Syntactic uniformity does not imply an absence of lexical categories. As Himmelmann (2008: 249) puts it, "all Tagalog content words (both roots and derived words) may all occur in essentially the same basic syntactic positions", but with regard to lexical categories "there are clear-cut categorial distinctions".

Still, the occurrence of verbs in DPs is a challenge for many syntactic models in which lexical categories are directly associated with syntactic functions. As a workaround, a verb combined with a determiner (as in (2)) is sometimes analysed as a nominalization or as a headless relative clause (Harris & Campbell 1995: 153). Indeed, assuming a zero-marked relativization or nominalization of the verb in a DP can be useful when comparing these constructions to their functional counterparts in other languages, such as the English cleft construction by which (2) can be translated. At the same time, it is less easy to distinguish relative clauses or nominalizations from simple noun phrases than in languages in which the difference is overtly morphologically marked. More importantly, such an analysis blurs an important grammatical property of the languages in question, which is that no derivation or other marking is needed for a verb to be used in a referential expression.

Himmelmann (2008), too, expresses an uneasiness regarding a relativization analysis of what I call here "verbal DPs" (i.e., DPs containing a morphological verb) in Tagalog. One of his arguments reads as follows:

> [F]rom a crosslinguistic point of view, any analysis of Tagalog clause structure should be able to express the fact that ACTION-words can be used in the same phrase-structural positions as OBJECT-words without any extra morphological marking and the fact that such use is clearly more common in Tagalog than in languages which require nominalisation or relative clause formation in order to achieve the same functional effects. (Himmelmann 2008: 268)

This statement can be divided into three separate components: (i) the finding that in some languages, verbs can productively be used in DPs without further marking, while most languages require some overt morphosyntactic operation; (ii) the claim that verbs in DPs are more common in Tagalog than headless relative clauses or participant-oriented nominalizations in other languages; and (iii) that there are certain "functional effects" associated with the placement of a verb in a DP or, in other languages, with the use of a relative clause or of a participant-oriented nominalization.

The aim of this paper is to explore in more detail the "functional effects" of the placement of a verb in a DP, which Himmelmann (2008) does not elaborate on. I will investigate this on the basis of a corpus of spontaneous oral discourse in Movima, a linguistic isolate of lowland Bolivia, whose structural properties are very similar to those of Tagalog (Haude 2025).[1] Movima content words can be classed as nouns, verbs, and adjectives; and all content words can directly function as predicates or be productively placed in a DP without any morphological marking. The following are the main characteristics of the system:

- There is no copula; all verbs, unpossessed common nouns, and adjectives can function as predicates, sharing the same syntactic properties (e.g. regarding the ways in which their arguments can be expressed).
- All content words can occur in a DP.
- DPs containing an underived content word are participant-oriented: they refer to the participant that would be the syntactically privileged argument of the content word in predicate function (i.e. the single argument of nouns and intransitive verbs and the external argument of transitive verbs).

The pattern is illustrated in (3). (If not indicated otherwise, all examples in this paper are from the corpus of spontaneous Movima discourse on which this study is based; see below.) In (3a), which is a prototypical transitive clause, the predicate

1. At the time the data were collected, Movima was spoken by a few hundred elderly speakers in Santa Ana del Yacuma, Bolivia. It is an endangered language, and the corpus was compiled during the DOBES documentation project "Documenting Movima, a linguistic isolate of the Moxos region (Bolivia)" (2006–2013). Most of the corpus is archived at The Language Archive and is accessible upon request.

is a transitive verb, and the DP expressing the P argument — the external argument — contains a noun.[2] In (3b), by contrast, the predicate is a noun — the clause is therefore intransitive — and the argument DP contains a transitive verb. The A argument is encoded like a possessor (see below), so that the construction can adequately be paraphrased as an English equational clause with a possessed DP. (See Himmelmann (1991) and Kaufman (2009) for a similar analysis of Tagalog.) A more idiomatic English translation involves a headless relative clause, which signals that the referent of the DP is only a participant in the "specific instance of the action denoted by the root" (Himmelmann 2008: 287).[3]

(3) a. ***yey-na***=*us* *[is* *mo'incho]*
want-DR=3M.AB ART.PL *chivé*
'He wanted *chivé* (fermented manioc).' [LTC_020906_5 381]

b. ***to:mi*** *[os* *yey-na=us]*
water ART.N.PST want-DR=3M.AB
'Water (was what) he wanted.' (Lit. "Water was his wanted (thing).")
[EAO_240807_vibora 133]

The structural properties of Movima verbs and nouns in predicate vs. argument function are discussed in Section 2. The criteria that distinguish these word classes are presented in 2.1, and the properties of DPs containing verbs ("verbal DPs") are outlined in 2.2; they include the participant orientation of DPs (2.2.1) and the syntactic operations that the content word inside a DP can undergo, which are a valency-decreasing operation and negation (2.2.2).[4]

Section 3 turns to the functional effects of verbal DPs in argument function. Corpus data show that, interestingly, the distribution of verbal DPs is opposed to that of the typical, nominal DPs (i.e. DPs containing a noun): verbal DPs are more

2. A note on terminology: "A" and "P" are used for the most agent-like and the most patient-like syntactic argument of a transitive predicate. "S" refers to the single argument of an intransitive clause. "Actor" and "undergoer" designate the role of an event participant, independently of the verb's transitivity. The terms "internal argument" and "external argument" are based on constituency and independent of semantic role. This is due to the Movima direct-inverse-system, which is comparable to Tagalog voice (Haude 2025); the fact that the internal argument of direct-marked verbs is A corresponds to the ergative structure of clauses with direct-marked predicates. The external argument, which encodes P of direct-marked predicates, has syntactic subject properties (Haude 2019a).

3. In the examples, argument DPs are marked by square brackets, and predicates are given in boldface. Tense and definiteness in the English translations — categories that are not systematically encoded in Movima — correspond to the example's context.

4. For more information on the phenomena described here, the reader is referred to the Movima grammar (Haude 2006), as well as to other sources as indicated in the text.

common with nonverbal predicates (nouns, quantifiers, and demonstratives), in contrast to the usual pattern in which the most common predicates are verbs (3.1). This points to the pragmatically marked status of verbal DPs, but the effects are different according to the predicate the verbal DP occurs with. When the predicate of a clause with a verbal argument DP is a noun, then the effect is that of participant focus (3.2). When the verbal DP occurs with a predicate that is a quantifier (3.3) or a demonstrative (3.4), i.e. in a quantificational or existential clause, the effect is rather that of nonspecific or indefinite reference: the referents are only characterized by their participation in a state or event, not by their inherent properties. Section 4 summarizes the findings and concludes that in fact, the functional effects of verbal DPs in Movima indeed correspond to those of relative clauses in other languages. However, verbal DPs achieve these functions in a far more economical way than relative clauses, since they do not involve any additional marking (or "referential scaffolding", Kaufman 2025).

2. Syntactic uniformity in Movima

2.1 Structural similarities and differences between nominal and verbal predicates

In order to speak about the syntactic uniformity of lexical categories in a language, it must first be established on what level and to what degree the language distinguishes between these categories. In Movima, there is no inflectional marking that distinguishes between nouns and verbs, such as morphological number marking on nouns or tense-aspect-mood marking on verbs. Rather, lexical categories can be established on the basis of the combinability of lexical bases with derivational morphemes. So, for instance, noun roots (and sometimes also full nouns) can be incorporated into verbs, but verbs or verb roots cannot. Nouns — and also adjectives (i.e. property denoting words) — can be combined with a dynamic verbalizing suffix ('to make/do something with N'; see (15)), but verbs cannot (see Haude 2006: 485–489). Two good rules of thumb for distinguishing between nouns, verbs and adjectives are given in (4); note that they both involve the morphosyntactic properties of full words, be they morphologically basic or derived.[5]

5. Within the thus established categories, the most important lexical subcategories are transitive vs. intransitive verbs, possessed vs. unpossessed nouns, and property-denoting adjectives vs. quantifiers.

(4) Working criteria for distinguishing nouns, verbs, and adjectives in Movima
   i. Person encoding: in principle, all nouns can be combined with an internal enclitic (explained below) that encodes a possessor. Verbs, in contrast, can only take an internal enclitic if they are overtly morphologically marked as transitive by a 'direct' or 'inverse' marker. Adjectives cannot be combined with an internal enclitic at all.
   ii. Marking in subordination: As predicate of a complement/adverbial/ negated clause, nouns and adjectives are reduplicated or receive a suffix *-łe*. Verbs, in contrast, receive a suffix *-wa* and cannot be reduplicated. Nouns combined with the suffix *-wa* acquire a dynamic reading.

(Haude 2011)

Section 2.1.1 describes the differences in person-marking patterns of verbs and nouns. Section 2.1.2 presents the predicate function of different lexical classes and illustrates the formal differences of nominal and verbal predicates in subordinate (i.e. complement and adverbial) clauses.

2.1.1 *Identical person encoding on possessed nouns and transitive verbs*

According to criterion i., any noun can in principle be marked as possessed. In (5), the words *kweya* 'woman' and *alwaj* 'spouse' (cooccurring appositionally in one single DP) are marked as possessed. The possessor is encoded by an "internally" encliticized pronoun, leading to a new prosodic word with stress on the new penultimate syllable and to the insertion of an epenthetic vowel after a consonant-final host, as on *alwaj* in (5). (Throughout this section, the relevant stress is represented by the accent symbol.)

(5) *ban jayna* ***joy-cheł*** *[isnos kweyá=us, alwaj-á=us]*
but DSC go-R/R ART.F.PST woman=3M.AB spouse-EP=3M.AB
'But his woman, his wife, had already left.' [EAO_In between 198]

Nominal possessor encoding is identical to the encoding of the internal argument of a transitive verb. However, in contrast to nouns, transitive verbs are always overtly marked: direct marking indicates that the internal argument encodes the actor, as in (6a), and inverse marking indicates that the internal argument encodes the undergoer, as in (6b). As in (5) above, the penultimate stress is indicated by an accent; an epenthetic vowel can be observed on the inverse-marked verb, whose base ends in the consonant *y*.[6] Internal cliticization applies to pronouns, as in the examples given here, but also to the determiner when the possessor or the internal argument is expressed by a full DP (see e.g. (30) below).

6. In Movima, /w/ and /y/ are consonants. Note that the orthographic symbol *y* represents /j/, while the symbol *j* represents /h/, as in Spanish.

(6) a. direct
man<a>yé=us *[os mi:chi]*
find<DR>=3M.AB ART.N.PST cat
'He found a cat.' [JGD_130907-13 043]

b. inverse
lap-kay-á=us *[os mimi:di]*
bit-INV-EP=3M.AB ART.N.PST snake
'He was bitten by a snake.' [EAO_240807_vibora 006]

Transitive verbs that are not combined with an overt internal enclitic are obligatorily interpreted as having a first-person singular argument, as shown in (7). This is the same with inalienably possessed nouns, like *alwaj* 'my spouse' in (8). Therefore, on a transitive verb or an inalienably possessed noun, the first person singular is represented as a zero enclitic, =Ø.

(7) ***man<a:>ye=Ø*** *[is rey oy-ra di' lari:ya]*
find<DR>=1SG ART.PL EPIST two-CLF.NTR REL brick
'I found two bricks.' (Lit. "... two things that (were) bricks.") [EAO Sueño 069]

(8) *jayna t* ***joy-cheɬ*** *n-us jayna alwaj=Ø*
DSC 1INTR go-R/R OBL-ART.M DSC spouse=1SG
'Then I left with my new husband.' [JAO Naye 063]

A verb without a direct or inverse marker is intransitive and cannot take an internal enclitic. When the single argument of an intransitive verb is encoded by a bound pronoun, this pronoun is attached through "external" cliticization, whose prosodic effects are different from that of internal cliticization. External cliticization is represented by two hyphens (--). Stress remains on its usual penultimate place of the verb base, as shown by the accent symbol in (9), and rather than requiring a preceding vowel, an external enclitic resyllabifies with a host-final consonant. External cliticization applies to the same pronominal forms as possession or internal argument encoding. It does not apply to determiners.

(9) *jayna* ***jóy-cheɬ****--is*
DSC go-R/R--3PL.AB
'Then they left.' [Escape Marivel 075]

Based on criterion i., some words that would intuitively be considered verbs must be classed as nouns. This is the case, for instance, with *jankwa* 'say' (lit. "said (thing)"), which is obligatorily combined with an internal enclitic but does not contain a morphological marker of transitivity (see (32) below).

Adjectives cannot be marked as possessed, which is why they are not easily distinguished from intransitive verbs. Here, criterion ii. is needed, which involves

the marking of predicates in subordinate clauses. Criterion ii. is illustrated in the following subsection, which deals with the predicative use of content words.

2.1.2 *Verbal and nominal predicates*

Movima has no copula in basic affirmative clauses, and unpossessed common nouns can function directly as predicates of intransitive clauses.[7] As in a verbal intransitive clause, the argument of a nominal predicate can be expressed as a DP (10), as an enclitic pronoun (11), or it may remain unexpressed (12).

(10) ***itila:kwa*** *[os majni=Ø]*
man ART.N.PST offspring=1SG
'My son was a man.' [EAO Cbba 110]

(11) *jayna* ***itila:kwa****--us*
already man--3M.AB
'He was already a man.' [JZH_080807 201]

(12) *pekłełe* ***itila:kwa***
all man
'(They were) all men.' [EAO_120906_1 223]

While nonpossessed common nouns can freely occur as predicates, the possibilities are much more limited for possessed or proper nouns (see Haude 2019b). This is probably due to their entailing specific reference (they usually have an identifiable referent), while predication ascribes a property or class membership to an entity. Identificational clauses, in which two specific referents are equated, are formed with a different construction in Movima, which includes a pronominal copula and will not be discussed here (but see Haude 2018).

Verbal and nominal predicates can be distinguished by the form they take in adverbial and complement (including negated) clauses. Here, verbal predicates receive the suffix *-wa* (glossed as 'event nominalizer') and are marked as possessed, as in (13). Unlike DPs with underived verbs, which refer to an event participant (see 2.2.1), DPs containing a form derived with *-wa* refer to an event or non-time-stable state, as in (13).

(13) ***sotak-sema:na*** *[oł terani:-****wa****=Ø]*
one-week ART.PST.N:1 ill-NMLZ.EVT=1SG
'I was ill for one week.' (Lit. "My being ill (was) one week.")
[ERM_150806 674]

7. Free pronouns are syntactic predicates that can be analyzed as a copula. However, they form non-basic clauses (Haude 2018), and they cannot link two clausal constituents.

In contrast to verbs, nominal predicates of complement and adverbial clauses either undergo partial reduplication or receive the suffix *-łe* (apparently in free alternation), and the derived form is marked as possessed.[8] The derived form is preceded by a determiner and can be analyzed as a nominalization denoting the property of "being X" (see Haude 2011). Example (14) shows a nominal predicate in an adverbial clause.

(14) *n-os* *jayna itila<**kwa:**~>kwa=Ø*
OBL-ART.N.PST DSC man<NMLZ.ST~>=1SG
'When I was a man already …' (lit. "At my already being a man …")
[HRR_081009_isbijaw 268]

Combining the criteria in (i) (person marking) and (ii) (nominalization pattern) helps to identify a third lexical category, which contains property- and quantity-denoting words that can therefore be labelled adjectives. These words cannot be marked as possessed (unlike nouns), but they undergo reduplication when forming subordinate predicates (like nouns). Consider the word *kawra* 'a lot' in (15). (The syllable-initial /r/ of the reduplicated syllable surfaces as [d].)

(15) *jayna n-os* *kaw-<**ra**~>da=is* *lopa:vos jayna, jayna*
DSC OBL-ART.N.PST much-<NMLZ.ST~>CLF.NTR=ART.PL manioc DSC DSC
***mo'incho:-tik**--iy'łi jayna*
chivé-VBZ.EVT--1PL DSC
'Then, when there was a lot of manioc (lit. "at the manioc's being much"), we made *chivé* (fermented manioc flour).' [EAO_Chaco III 015]

Instead of being reduplicated, adjectives — and, although this is less common, also nouns — can also receive the suffix *-łe*. It is not clear whether the difference is systematic: it seems that both forms are equally acceptable.

Some words that might be classed as intransitive verbs because they seem to denote a non-time-stable concept, such as *bispa* 'know', are reduplicated in subordination but never found in a possessive construction; consequently, these words belong formally to the class of adjectives.

(16) *ka=s* *bis<**pa**~>pa=us* *n-as*
COP.NEG=DET know<NMLZ.ST~>=3M.AB OBL-ART.N
*ben-a-bet-**wa**=us*
draw-DR-BR.hide-NMLZ.EVT=3M.AB
'He didn't know how to write.' (Lit. "His being knowledgeable in his writing was not.") [EAO_Abuelo 002]

8. Nominal bases can take the verbal suffix *-wa*, too, which then adds a dynamic reading; see Haude (2011).

Some words that are nouns according to criterion (i) are nominalized by a suffix *-niwa*, which is a fossilized combination of a verbalizing suffix *-ni* (elsewhere, a verbalizer of 'process') and the verbal nominalizer *-wa*. This is the case of *jankwa* 'say', shown in the adverbial clause in (17). Also free personal pronouns and demonstratives are nominalized in this way when functioning as predicates (see (22) below).

(17) *jayna* ***mambaycho*** *n-as* *jayna jankwa-**niwa**=is* *dotor*
DSC satisfied OBL-ART.N DSC say-VBLZ:NMLZ=ART.PL doctor
"di:ra=s ***ya:yu'u"***
at_least=2INTR be_fine
'(I am) already satisfied when the doctors say "You're okay".'
[ERM_140806_1 0184]

Hence, the morphological derivation that predicates of adverbial and complement clauses undergo is a way to distinguish predicates of different lexical categories. Since adverbial and complement (including negated) clauses are very frequent in discourse, it is usually possible to find corpus evidence of the lexical category a lexeme belongs to. At the same time, since nouns, adjectives, and intransitive verbs can only be distinguished indirectly through tests like these, there will always be words whose category remains undetermined.

2.2 The distribution of verbal DPs in discourse

2.2.1 *Orientation: Verbal DPs have a predictable meaning*

A canonical Movima DP consists minimally of a determiner and a content word.[9] The determiner is usually an "article" (neutral with regard to definiteness). The content word is a noun, a verb, or an adjective. A verbal DP refers to the argument that would be the external argument of the same verb in predicate function. The participant role of the referent is specified by the morphological and lexical properties of the verb. A DP containing a direct-marked transitive verb refers to the undergoer of the activity encoded by the verb, as in (18a) (see also (3b) above); a DP containing an inverse-marked transitive verb refers to the actor, as in (18b). (Note that boldface continues to mark main-clause predicates, while DPs encoding the external argument are enclosed in brackets.)

(18) a. ***ko'*** *łat [os* *joy-**a**-łe=us]*
tree;wood EV ART.N.PST go-DR-CO=3M.AB
'A tree trunk was what he brought.' [PMP_HRR_etal_210908 244]

9. In addition, DPs may also contain relative clauses and adjuncts.

b. ***ji<wa:~>wa*** *[os rey joy-łe-**kay**-a=us] ney*
come<MD~> ART.N.PST EPIST go-CO-INV-EP=3M.AB here
'Someone came here who took him (there).' (Lit. "His taker came.")[10]
[CCT_120907_1 100]

When the DP contains an intransitive verb, the referent is the participant that is encoded as the S argument of the verb in predicate function. In (19), the DP refers to the single participant of the event of "coming", which is an actor (cf. (18b), where this verb is the predicate). In (20b), the DP refers to the undergoer of a semantically bivalent verb that contains no direct or inverse marker. Such bivalent verbal bases without any voice marking are interpreted as undergoer-oriented, i.e. the single argument of a predicate of this type is the undergoer, as shown in (20a) (see Haude 2012).

(19) *ban **so:t-e** [os ji<wa:~>wa nokodé=s sot-di]*
but other-CLF.person ART.N.PST come<MD~> over_there=DET other-CLF.house
'But somebody else (was who) came over there to the other house.'
[EAO_Visita 024]

(20) a. ***rim-łe** [is li:mo]*
sell-CO ART.PL lemon
'The lemons are sold.' [ERM_150806 714]
b. ***o:kaka:-ra** [os rim-łe]*
all_kinds-CLF.NTR ART.N.PST sell-CO
'All kinds of things (is what is) sold.' [ATL_230806 030]

While the DP use of verbs is fully productive (250 different verb lexemes of different morphological types are attested in the corpus), there are some exceptions to the predictable participant orientation. These involve some intransitive verbs which, when placed in a DP, take (or can take) an internal argument enclitic, so that they appear as a noun (because only nouns can take possessor enclitics without direct or inverse marking). A case in point is the verb *ya:lo:we* 'drink'. As main-clause predicate, this verb is intransitive and actor oriented, i.e. the actor is encoded as the single argument, and the undergoer, if overtly expressed, is marked as oblique. This is shown in (21a). In a DP, this verb can retain this pattern, as in (21b): here, the DP that contains the verb *ya:lo:we* refers to the people who drink. However, as shown in (21c), the verb *ya:lo:we* can also be marked as possessed (here, by the second-person internal enclitic *=n*), and the DP then refers to the undergoer.

10. The neuter article (here, the past-tense form *os*), usually reserved for non-human referents, can be used for nonspecific reference to humans; this means that despite being encoded as "possessed", a verbal DP does not necessarily have a specific or definite referent.

(21) a. *jayna* ***ya:lo:we****--is* *n-is* *pokso*
DSC drink--3PL.AB OBL-ART.PL *chicha*
'Then they drank *chicha* (manioc beer).' [HRR_120808-tigregente 543]

b. ***kiro'*** *[kis* *ya:lo:we ni-kis* *da'ra-m-mes]*
DEM.PL.AB ART.PL.AB drink OBL-ART.PL.AB ostrich-LN-BR.grease
'There are (people who) drink ostrich grease (as medicine).'
[ERM_140806_1 0932]

c. ***to:mi*** *ja' [kis* *yá:lowe=n]*
water just ART.PL.AB drink=2
'Only water (was what) you drank?'[11] [ERM_150806 704]

Another case where the orientation of a verbal DP is different from the orientation of the verb in predicate function involves argument-incorporating verbs (see Haude 2019a).

Since verbs can only be combined with an internal enclitic when they are overtly marked as transitive, while nouns can receive an internal enclitic independently of their form, intransitive verbs that take an internal enclitic inside a DP must be considered nominal forms. They are lexical exceptions from the rule according to which all Movima DPs refer to the concept that would be encoded as the external (or single) argument of the same content word in predicate function.[12]

2.2.2 *DP-internal operations: Valency decrease and negation*

The syntactic uniformity of verbs and nouns extends also to their syntactic properties inside a DP: both verbs and nouns in a DP can undergo a valency decreasing operation, and both verbs and nouns can be negated.[13] In discourse, however, these processes are much more frequent with verbs than with nouns, showing that despite their syntactic uniformity, the difference between entity- and event-denoting expressions does have an influence on their use.

The valency decrease works as follows: the content word in the DP is preceded by a particle *kaw* (or its free variant *kwey*), and it cannot take an internal enclitic. The valency-decreased DP refers to the possessor of the noun (as in (22)) or to the entity that would be encoded as the internal argument of the verb (as

---

11. The question is marked prosodically, with stress and high pitch on the first syllable of the predicate (Haude 2006: 54–55).

12. Another example of this kind of conversion is *chonsineł* 'native language; speak the native language' (Haude 2009).

13. The operations described here are not possible with main-clause predicates of any kind. They are restricted to constructions that can be analysed in terms of extraction, i.e. topicalization and relative-clause formation (Haude 2019a).

in (23)).[14] The operation is not possible with inherently monovalent words, i.e. intransitive verbs or adjectives. When the possessed entity and the external argument are overtly expressed, they are marked as oblique, as in (22) (with a noun) and (23) (with a direct-marked verb).

(22) *bo* ***ka****=s* *rey* *oso'-****niwa*** *[os* ***kaw*** *rey*
REAS COP.NEG=DET MOD DEM.N.PST-VBLZ:NMLZ ART.N.PST VALDEC EPIST
no:no *n-isko]*
domestic_animal OBL-PRO.PL.AB
'Because they (the cows) did not have an owner.' (Lit. "There was not an animal-owner of/for them.") [GBM Ganado 004]

(23) *noso'* ***iso'*** *[is* ***kwey*** *ji:sa:-na n-is* *bu~but-kwa]*
at_that DEM.PL.PST ART.PL VALDEC make-DR OBL-ART.PL RED~BR.mud-ABS
'In those (times) there were (people who) made clay (for pottery).'
[CCT_120907_2 042]

Elicitation shows that the valency-decreasing operation is fully productive. However, with nouns the construction is often replaced by a lexicalized complex expression *kaw n-inła* 'owner (lit. VALDEC OBL-PRO.1SG)' (Haude 2006: 537). This shows that while syntactically, (possessed) nouns and (transitive) verbs can undergo the same syntactic operation, in actual use they are often treated differently.

This pattern is also observed in the second, and for the present discussion more important, syntactic parallel between verbs and nouns in DPs: both verbs and nouns inside a DP can be negated, but especially with nouns, this process is extremely rare. When the content word in a DP is negated, a particle *loy* is inserted before the content word.[15] Intransitive verbs are then marked by the suffix *-wa*, as in (24), and nouns are either reduplicated or receive the suffix *-łe*, as in (25), which is reminiscent of the nominalization pattern in adverbial and complement clauses (and is therefore glossed in the same way). Transitive verbs, however, do not receive any further marking, as shown by the boldface form in (26) (which heads a relative clause with the same pattern).

(24) *jayna* ***iso'*** *[is* ***loy*** *jaywaneł-****wa]***
DSC DEM.PL.PST ART.PL NEG.SUB agree-NMLZ.EVT
'Then there were those/some who did not agree.'
[JGD_160808-Fundacion_2 340]

14. This operation is most common with direct-marked verbs, where it functions as an antipassive, and almost never occurs with inverse-marked verbs (Haude 2024).

15. The particle *loy* is also used to negate subordinate clauses (Haude 2006: 544–545).

(25) ***ka***=*s* *rey* *ja'* *sal-na-wa=i* *[kos* ***loy*** *rey*
COP.NEG=DET EPIST just look_for-DR-NMLZ.EVT=3PL ART.N.AB NEG.SUB EPIST
*mowimaj-**łe**]*
Movima-NMLZ.ST
'They don't just look for someone who is not Movima.' (I.e., they look for someone who *is* Movima.) [EAO Tolkosya II 014]

(26) *bo* ***oso'*** *[os* ***loy*** *rey* *ten<a>pan**ł**e=i* *di'* *loy*
REAS DEM.N.PST ART.N.PST NEG.SUB EPIST be_able_to<DR>=3PL REL NEG.SUB
rey *im<a>ra=i]*
EPIST imitate<DR>=3PL
'Because there was (something that) they could not (do), which they did not imitate.' [PMP_HRR_etal_210908 288]

The fact that DP-internal negation is productive with both verbs and nouns can be taken as evidence that the content word of any Movima DP (except perhaps proper names and possessed nouns) can be analysed as an embedded predicate (see Haspelmath 2016; Creissels 2021), in line with Launey's (1994, 2004) proposal of "omnipredicativity". As a consequence, all DPs can be analysed as light-headed relative clauses (the determiner being the non-lexical head that establishes a referent, whose identity is further specified by the content word).[16] However, DPs with a negated noun are hardly ever found in the corpus and are difficult to elicit, and there is no corpus example of a negated possessed noun.[17] Again, this seems to indicate that in actual use, a difference is made between entity- and event-denoting expressions, even though the grammar does not make this distinction.

Summing up, Movima verbs can occur in a DP without any morphological modification, and nouns and verbs can undergo the same syntactic operations inside a DP as verbs. Therefore, Movima can be regarded as a language with syntactic uniformity like Tagalog according to Himmelmann (2008). There are, however, differences between verbs and nouns in terms of the discourse frequency of these DP-internal operations: the valency decrease of possessed nouns is often replaced by a more lexicalized expression, and nouns in DPs are almost never negated. The following section goes in a similar direction by showing the untypical distribution of verbal DPs in discourse.

16. The term "light-headed relative clause" was introduced by Citko (2004).

17. However, elicitation shows that the negation of nouns, including possessed ones, is productive. In a recent telephone elicitation session (August 2025), a speaker offered a number of negated nouns inside DPs when provided with a fitting context ('I don't talk to (someone who) is not a woman, (someone who) is not my husband, etc.').

3. The discourse use of verbal DPs

3.1 The combination of verbal argument DPs with different predicate types

As in all languages (see Croft 2003: 185), in Movima the typical referential phrase (i.e., DP) contains a noun and the typical predicate is a verb. In a count of four Movima texts of altogether approximately 800 main and subordinate clauses, 80% of the predicates are verbs; only 11% are nouns, and the remaining 9% are adjectives or demonstratives.[18] These proportions are represented in Figure 1 (based on numbers given in Table 1 below).

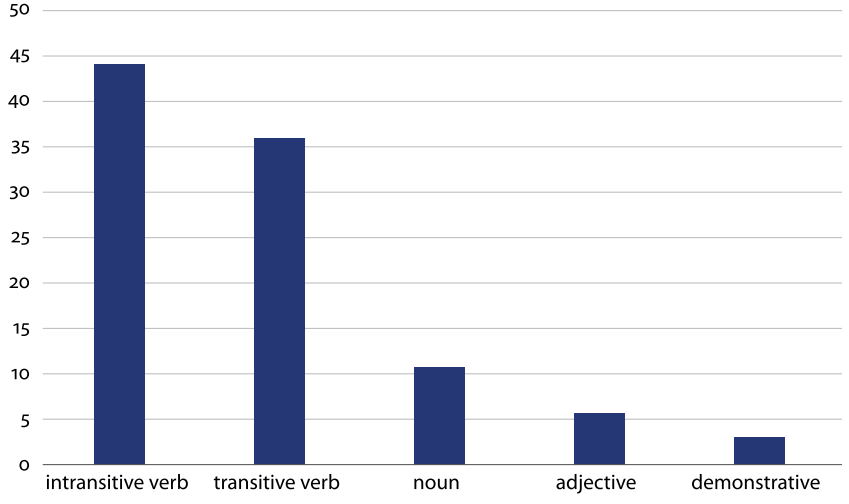


Figure 1. Percentages of different predicate types in four texts (800 clauses)

In these texts there are only 19 DPs containing a verb, in contrast to 780 DPs containing a noun.[19] The numbers — on which Figure 1 is based — are presented in Table 1 (the split of adjectives into quantifiers and property-denoting words will be explained below). The number of verbal DPs in this sample is so low that it does not even significantly affect the percentages, given in the bottom line.

The fact that verbal DPs are so extremely infrequent is a sign that their use is pragmatically marked. In order to get an idea of their function, I investigated with which predicate types verbal argument DPs co-occur in the overall corpus. This corpus, collected from about 50 female and male speakers during fieldwork

18. These texts were annotated according to the GRAID annotation schema (Haig & Schnell 2014), which allows a straightforward identification of predicate types.

19. In GRAID I annotated them as headless relative clauses, sometimes based on intuition.

Table 1. Predicate types in four texts (800 clauses; DPN=nominal DP, DPV=verbal DP, qtf=quantifier, prop=property denoting word)

Predicate:	DPN	DPV	Total	Proportion
intransitive verb	344	8	352	44%
transitive verb	282	3	285	36%
noun	85	4	89	11%
adjective (qtf)	32	1	33	4%
adjective (prop)	13	0	13	2%
demonstrative	24	3	27	3%
Total	**780**	**19**	**799**	**100%**

between 2001 and 2012, consists of approximately 30 hours of audio-recorded speech of mostly narrative monologic discourse, i.e. over 130,000 transcribed words and over 25,000 clauses. All DPs containing verbs — identifiable through the criteria given in Section 2.1 — were tagged manually. There are about 700 of them, of which about 500 occur in in argument function.[20] For each of these tokens I annotated the predicate type it occurs with.

The result is reproduced in Table 2, graphically represented in Figure 2. Here we see that verbal argument DPs, already rare, tend to co-occur with predicates that are overall infrequent, too: namely, nonverbal predicates.

Table 2. Predicate types occurring with verbal argument DPs (DPV) in the corpus

Predicate:	DPV	Proportion
intransitive verb	52	10%
transitive verb	92	18%
noun	109	21%
adjective (quantifier)	112	22%
adjective (property)	15	3%
demonstrative	129	25%
Total	**509**	**100%**

20. The 200 remaining non-argument DPs are either oblique-marked or left/right dislocated. Note that the number of verbal DPs is approximative because some highly frequent verbal DPs, like 'the so-called (one)', were not tagged every time.

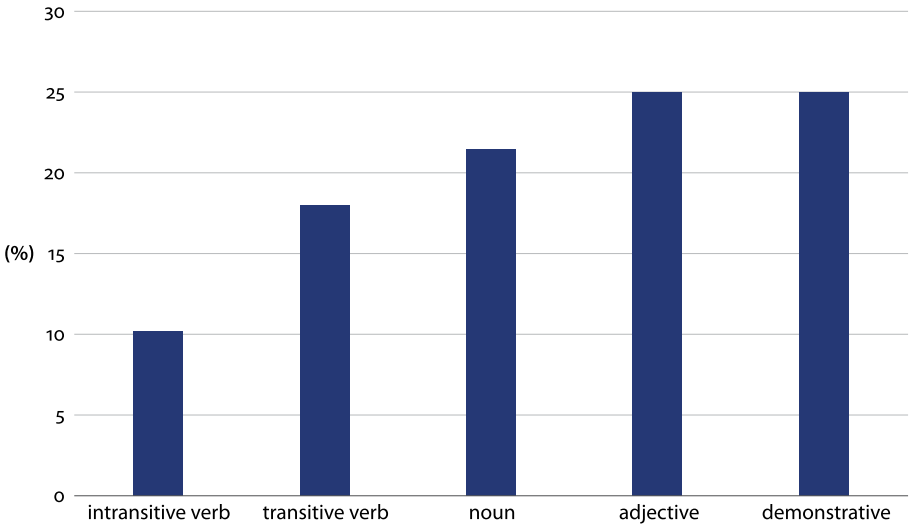


Figure 2. Distribution of verbal argument DPs ($n = 509$) with different predicate types in the corpus

When the distribution of predicate types in combination with verbal DPs is compared to the default distribution of predicate types shown in Figure 1 above, it becomes apparent that the former is almost exactly anti-proportional to the latter: verbal DPs occur most often with nonverbal, i.e. untypical predicates. (I have not investigated why verbal DPs occur more often with transitive verbal predicates, which, according to Figure 1 and Table 1, are normally less frequent than intransitive verbal predicates.)

This finding confirms what was proposed by Launey (2004) as an explanation of his concept of omnipredicativity: in an omnipredicative system, argument (or rather: participant) focus is reached by simply placing the noun in predicate position and the verb inside the argument expression. That this is indeed the case in Movima is shown in Section 3.2. Table 2 also shows that verbal DPs occur even more often with nonverbal predicates that are not nouns: in 25% of all clauses with verbal argument DPs, the predicate is an adjective, and in another 25% the predicate is a demonstrative; furthermore, of the adjectives, it is especially quantifiers that form the predicate of a clause with a verbal DP (22%), while property-denoting adjectives are remarkably rare (3%). Verbal DPs with quantifiers are discussed in 3.3, and verbal DPs with demonstrative predicates — which form existential clauses — are discussed in 3.4.

3.2 Nominal predicate, verbal argument: Focus

The combination of a verbal DP with a nominal predicate forms a participant focus construction (Haude 2021). The translation of Example (3b) with an English cleft already indicated the focus reading of a clause in which noun and verb exchange their positions. Example (27), with an additional negated clause, shows that this construction can express contrastive focus.

(27) ***rulrul*** *[os tikoy-na=Ø], ka=s rey tikoy<ak>-na=Ø*
jaguar ART.N.PST kill-DR=1SG COP.NEG=DET EPIST kill<IRR>-DR=1SG
n-is juyeni.
OBL-ART.PL person
'What I killed (was a) jaguar, I didn't kill a human (lit. "there wasn't anyone I killed of people"), you know.'[21] [PMP_HRR_etal_210908 165]

The focus function of a nominal predicate combined with a verbal argument is often (though not always) also reflected prosodically. Figure 3 and Figure 4 represent the prosodic contours of the above Examples (3a) and (3b), respectively. Example (3a) ('He wanted *chivé*') is a canonical clause, with a verbal predicate and an argument DP containing a noun. This type of utterance answers the hypothetical question 'What happened?'. The predicate receives high pitch, and so does the noun in the argument DP.

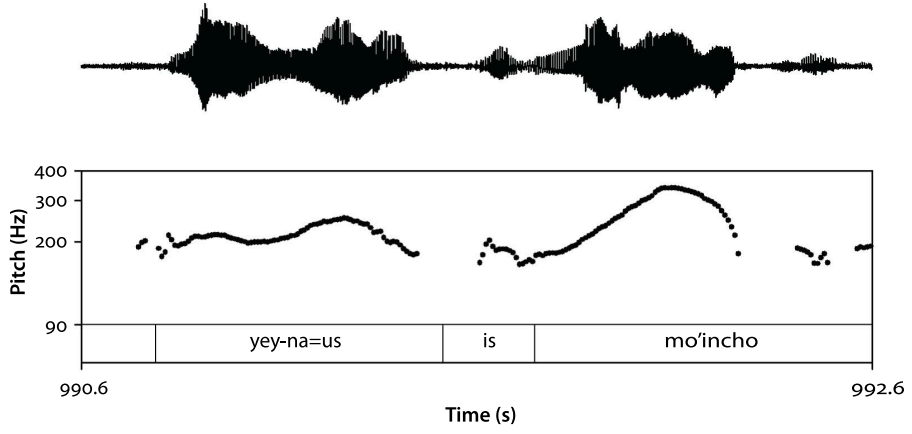


Figure 3. Prosodic contour of (3a) ('He wanted *chivé*.')

21. The "participant negation" with the irrealis infix still requires further analysis (Haude 2006: 439–442).

Figure 4 shows the prosodic contour of the focus construction with a nominal predicate and a verbal DP in (3b) ('Water was what he wanted.'). Here, the nominal predicate receives high pitch, while the prosodic prominence of the verbal DP (and the discourse particle *jayna* that follows it and is included here) is reduced. This utterance answers the hypothetical question 'What did he want?'.

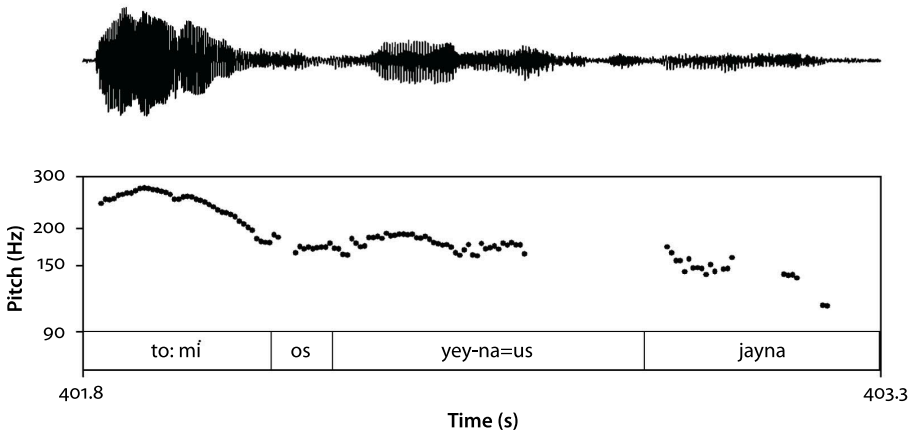


Figure 4. Prosodic contour of (3b) ('Water was what he wanted, then.')

The intonation patterns of Movima clauses in which the syntactic positions of noun and verb are inverted corroborates Launey's (1994; 2004) claim that in an "omnipredicative" system, the use of a noun in predicate function combined with a verb in argument function is a marked focus construction. The construction corresponds to an English "pseudo-cleft" (or "Wh-cleft"), i.e. an equational construction in which the clefted constituent is a nonverbal predicate (in (3b), '(it) was water') and the cleft clause is a light-headed relative clause ('what he wanted'). However, syntactic uniformity (or omnipredicativity) clearly is a more economic way to transmit pragmatically marked information.

3.3 Quantifiers as predicates

As can be seen from Figure 2 and Table 2, nominal predicates are by far not the most frequent nonverbal predicate types that verbal DPs cooccur with. Adjectives and demonstratives are more than twice as common in combination with verbal argument DPs than nouns. And among the adjectives, it is especially those that express notions like 'many', 'few', 'other', and numbers, which combine with verbal arguments. All these expressions have in common that they categorize entities in relation to other entities of the same type, and are hence subsumed

under the label "quantifier" in Table 2.[22] Examples of clauses with a quantifier as predicate were already given in (15) (with a nominal DP), (19), and (20b) (with verbal DPs).

According to the criteria given in Section 2.1, quantifiers belong to the class of adjectives: unlike nouns, they cannot be possessed, and unlike verbs, they are reduplicated when acting as subordinate predicates (as in (15)). A quantifier predicate (here, a number term borrowed from Spanish *treinta*) with a verbal argument DP can be seen in (28) (see also (19) and (20b) above).

(28) ***treynta*** *[is ji<wa:~>wa ney]*
thirty ART.PL come<MD~> here
'Thirty (were who) came here.' [CCT_120907_1 013]

Figure 5 shows that also in utterances with a quantifier as predicate, the highest prosodic prominence is on the main predicate, as in the focus construction with a nominal predicate (cf. Figure 4 above).

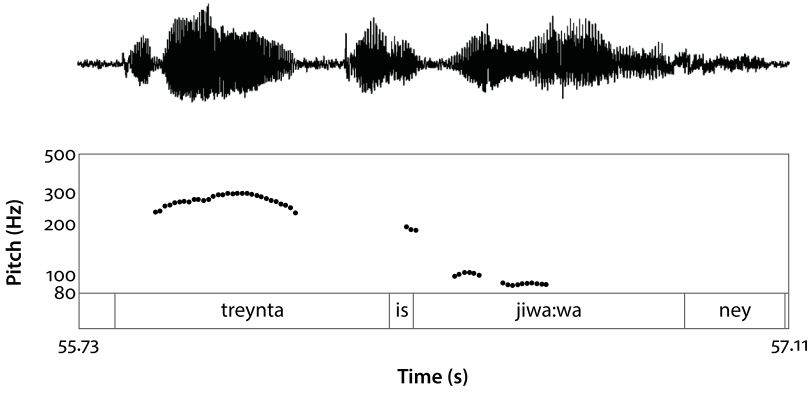


Figure 5. Prosodic contour of (28) ('Thirty (were who) came here.')

There must be an explanation of why quantifiers are so particularly common predicates of clauses with verbal DPs. It may have something to do with the fact that quantifiers do not say something about a referent in the way that verbal, nominal or property-denoting predicates do. Rather than characterizing an entity with respect to its properties or to the situation it is involved in, a clause with a quantifier as predicate presupposes the existence of an entity with respect to other entities (e.g., things that are sold in (20b) or people who came in (19) and (28)), either

22. The universal quantifiers *ba:ra* and *pekłełe* 'all' behave rather like particles than as predicates (see Haude 2019a) and are therefore not included here.

because these are included or because they are different from them. An explanation may, therefore, go along the lines of specific and nonspecific reference, but for the moment the question remains unsolved.

The domain in which verbal argument DPs are most common are existential constructions, treated in the following section.

3.4 Demonstratives as predicates (existential constructions)

In Movima, existential constructions are formed with a demonstrative pronoun as predicate (see Haude 2006: 295–296). Examples were already given in (21b), (22)–(24), and (26). An existential construction with a verb is illustrated in (29) (the preceding context is given in parentheses in the translation line).

(29) *che **kiro'*** *[kis joy-cheł]*
and DEM.PL.AB ART.PL.AB go-R/R
'(There are the (ones who) always come back,) and there are the (ones who) leave.' [ERM_140806_1 0680-0681]

That the demonstrative pronoun is the predicate of the existential construction is evident from the fact that it undergoes nominalization when the construction is subordinated, as in (22) above.

Existential (or "presentational") sentences are thetic constructions (Sasse 1987). Their main function is to introduce an entity, and they do not show the topic-comment partition of a categorical sentence. Accordingly, the demonstrative predicate does not receive the main prosodic weight as lexical predicates do. Figure 6 shows that it is the content word inside the DP (here, a verb) that receives highest prosodic prominence.

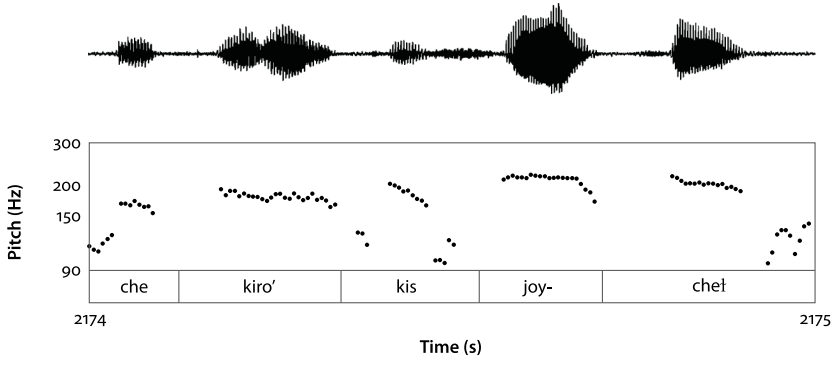


Figure 6. Prosodic contour of (29) ('and there are the (ones who) leave')

Demonstrative predicates can also express possession, as shown in (30). In this case, the DP is marked as possessed. The use of a demonstrative is an alternative to the reduplicated noun that forms a possessive predicate in Movima ('be provided with N') (Haude 2006: 296–297). In (30), the speaker uses both constructions: first the reduplication and then the existential construction with a possessed DP. The difference between the two constructions is that the argument of the reduplicated predicate is the possessor (in (30) *kos ney* 'the one here'), while the argument of the demonstrative predicate is the possessee.

(30) ***bari~ba:ri*** *[kos ney], eney,* ***koro'*** *[kos bari=os meche:ro], ney*
POSS~leg ART.N.AB here FILLER DEM.N.AB ART.N.AB leg=ART.N.PST oil_lamp here
'The (one) here had a leg, er, the oil lamp had a leg, here.' (Lit. "There was the oil lamp's leg here.") [CCT_120907_2 196]

Since specificity or definiteness are not grammatically encoded in Movima, the referent of the argument DP in the existential construction can be interpreted as either specific or non-specific, depending on the context. For instance, in (31), the inalienably possessed noun can be considered as having a unique referent. However, the sentence can also simply mean that the speaker was married.

(31) ***uso'*** *[us alwaj=Ø]*
DEM.M.PST ART.M spouse=1SG
'There was my husband'; or: 'I had a husband.' [ATL_230806 063]

When the existential construction contains a verb, as in (21b), (24), and (26) above, the situation is similar: often, the referent is non-specific (as in (21b) above 'There are people/those who drink ostrich grease'), but it may also be specific. In (32), the referent of the verbal DP can be interpreted as specific since it is further qualified by the predicate of the following clause (*jiranni* 'nice').

(32) *"****koro'*** *[kos eney* ***man<a>ye=y'ɬi*** *di' bet'i],* ***jiranni,"*** *jankwa=is*
DEM.N.AB ART.N.AB FILLER find<DR>=1PL REL grassland nice say=3PL.AB
'"We found a, er, (piece of) grassland (lit. "There was (what) we found, which is grassland"), (it) is nice", they said.' [JGD_160808-Fundacion_1 772]

Why do verbal DPs occur particularly often in existential constructions, as shown by the corpus count? Languages without syntactic uniformity, like English, use indefinite pronouns ('someone/something') or semantically light nouns ('people', 'things' etc.) followed by a relative clause in order to assert the existence of entities that are involved in states or events. Movima has no indefinite pronouns, and

there is no word for 'thing' that could form a semantically light lexical head of a relative clause; instead, a verbal DP is used.[23]

4. Conclusion: The predicate status of content words in DPs

This study has shown that in Movima, verbs, nouns, and adjectives display "syntactic uniformity" as defined by Himmelmann (2008): apart from some restrictions on possessed and proper nouns to function as predicates of basic clauses, there is no principled syntactic restriction on any class of content words to function as either predicates or arguments. Furthermore, it was shown that all types of content words inside a DP can be analyzed as a predicate: even a noun inside a DP can be negated. Crucially, no evidence was found that would identify nouns inside a DP as *not* being a predicate, since there are no systematic syntactic differences between nouns and verbs in DPs. Therefore, when an analysis in terms of relative clauses is considered, such an analysis must include *all* Movima DPs, also those containing a noun.[24]

One might want to argue that only verbal DPs should be analyzed as relative clauses, in line with semantic differences between nouns and verbs and as reflected by the translation into English or Spanish.[25] However, in contrast to languages where relative clauses are overtly marked or where the difference between verbs and nouns is straightforward, in Movima it would often be difficult to determine whether one is dealing with a relative clause or with a simple noun phrase: unless there is unambiguous overt marking, such as verbal voice morphology or a verbalizing suffix on a nominal base, the category of a word can only be determined indirectly through tests (see e.g. the example of *bispa* 'know' in (16)).

At the same time, it is clear that also in Movima there are semantic differences between entity-, event-, and property denoting words, and these differences are reflected by the pragmatic effect of placing a content word in an untypical syntactic position. As was shown above, nonverbal predicates and especially verbal DPs are rare in Movima spoken discourse, and verbal argument DPs occur particularly often in combination with nonverbal predicates.

---

23. The word *juyeni* 'person', by contrast, is quite common and occurs particularly often as part of a headed relative clause (either as head or as predicate); further research may show to what degree this construction is used to replace verbal DPs.

24. Alternatively, all content words in Movima DPs can be considered oriented nominalizations (cf. Kaufman 2009). I do not see any significant difference between the two analyses.

25. One reviewer observed that a relative-clause analysis of verbal DPs would be adequate because verbal DPs can include more clausal material, as in (23). However, this is not a sufficiently strong structural criterion since also nominal DPs may contain such material, as in (22).

Based on the findings of the present study, I propose that the use of verbal DPs in Movima has two major functions:

i. Backgrounding in a participant focus construction
   In accordance with Launey's (1994; 2004) account of what he calls "omnipredicativity", in Movima the predicate position is reserved for the focus part of the sentence, and the argument position hosts the sentence topic, i.e. the entity about which something is said. Placing a noun in predicate position has the effect of focusing on the event participant denoted by the noun, and placing the verb inside the argument DP has the effect of backgrounding the event description. This is comparable to the effect of cleft constructions in languages without syntactic uniformity, where the clefted constituent is expressed as a predicate ('N is') and the cleft clause as a headless relative clause ('what Vs') (Malcher 2021).
ii. Nonspecific reference
   Using a verbal DP permits to avoid characterizing the entity one is talking about. In languages like English this is done by relative clauses introduced by indefinite pronouns like "something/someone that/who" or "the one(s) that/who".[26]

Interestingly, like Movima, also Tagalog lacks indefinite pronouns: "'[s]omebody' and 'something' are expressed by an existential construction consisting of an existential particle plus a verb form" (Haspelmath 2013). In a Tagalog existential construction, however, the noun or verb directly follows the existential particle (Himmelmann 2008: 267), while in the Movima existential construction the argument DP always contains an article. This difference is perhaps related to the fact that the Tagalog determiner *ang* encodes specificity, while the Movima articles do not.

It seems like a promising line of research to compare the discourse use of verbal DPs in languages with syntactic uniformity with the use of comparable constructions in other languages. The parallels between the focalizing function of the "noun-verb-flip" (i) and clefts in other languages is obvious. Regarding the functions of verbal DPs that are related to nonspecific reference and indefiniteness (ii), it would be interesting to see in which syntactic environments headless relative clauses (in languages that have them), relative clauses headed by indefinite pronouns (in languages that have them), or relative clauses headed by light nouns tend to occur in discourse.

26. It should be noted that headed relative clauses, which can modify any kind of DP irrespective of lexical category, abound in Movima. A verbal DP specified by an additional relative clause was shown in (37).

To conclude, in terms of syntactic structure, all Movima DPs are identical, independently of whether they contain a noun or a verb (in correspondence with omnipredicativity in the sense of Launey 1994, 2004). The existence of a large corpus of spontaneous discourse data, however, has made it possible to add another dimension to the analysis, namely that of discourse distribution. Verbal DPs occur in marked syntactic environments, which are functionally parallel to clefts and to relative clauses with a semantically light lexical head. Therefore, comparing the distribution of Movima verbal DPs — where they can be identified as such — with light-headed or headless relative clauses in other languages can be fruitful for crosslinguistic comparison. Such a comparison should, however, bear in mind that Movima content words are syntactically uniform, and that the parallels with other languages take place primarily on a functional, not on a structural level.

Acknowledgments

I would like to thank the Movima speakers who generously shared their language, their stories and their experience with me. Further acknowledgments go to the two non-anonymous reviewers Eva van Lier and Eva Schultze-Bernd and to Uta Reinöhl, who fought their way through a previous version of this manuscript and provided highly insightful remarks. Thanks also to Katherine Walker for helping me with PRAAT. This paper is dedicated to Nikolaus Himmelmann, whose publications on Tagalog helped me understand how Movima works. I alone am responsible for all remaining mistakes.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following abbreviations are used:

--	external clitic	LK	linker
~	reduplication	LN	linking nasal
< >	infixation	MD	middle
=	internal clitic	NTR	neutral
AB	absential	OBV	obviative
ACT	actor	PRC	process
AV	actor voice	PRO	pronoun
BR	bound root	R/R	reflexive/reciprocal
CO	co-participant	REAS	reason
DR	direct	RED	reduplication
DSC	discontinuous	REL	relativizer
EP	epenthetic vowel	SPEC	specific article
EPIST	epistemic marker	ST	stative
EV	evidential	SUB	subordinate
EVT	event	VALDEC	valency decrease
INV	inverse	VBLZ	verbalizer.

References

Citko, Barbara. 2004. "On Headed, Headless, and Light-headed Relatives." *Natural Language and Linguistic Theory* 22 (1): 95–126.

Creissels, Denis. 2021. "Remarks on the Grammaticalization of Identificational Clefts." *Faits des Langues* 52 (1): 13–31.

Croft, William. 2003. *Typology and Universals*. 2nd edn. Cambridge: Cambridge University Press.

Haig, Geoffrey, and Stefan Schnell. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse) Manual Version 7.0*. Bamberg University.

Harris, Alice C., and Lyle Campbell. 1995. *Historical Syntax in Cross-Linguistic Perspective*. Cambridge and New York: Cambridge University Press.

Haspelmath, Martin. 2013. "Indefinite Pronouns." In *WALS Online (v2020.4)*, ed. by Matthew S. Dryer, and Martin Haspelmath. http://wals.info/chapter/46, Accessed on 2025-02-10.

Haspelmath, Martin. 2016. "The Serial Verb Construction: Comparative Concept and Cross-linguistic Generalizations." *Language and Linguistics* 17 (3): 291–319.

Haude, Katharina. 2006. A Grammar of Movima. PhD thesis, Radboud University Nijmegen. https://repository.ubn.ru.nl/bitstream/handle/2066/41395/41395.pdf

Haude, Katharina. 2009. "Reference and Predication in Movima." In *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*, ed. by Patience Epps, and Alexandre Arkhipov, 323–342. Berlin: De Gruyter Mouton.

Haude, Katharina. 2011. "Referring to States and Events: Subordination in Movima." In *Subordination in South American Languages*, ed. by Rik van Gijn, Katharina Haude, and Pieter Muysken, 141–168. Amsterdam: John Benjamins.

Haude, Katharina. 2012. "Undergoer Orientation in Movima." In *Ergativity, Valency and Voice*, ed. by Gilles Authier, and Katharina Haude, 159–287. Berlin and New York: De Gruyter Mouton.

Haude, Katharina. 2018. "Nonverbal Predication in Movima." In *Nonverbal Predication in Amazonian Languages*, ed. by Simon E. Overall, Rosa Vallejos, and Spike Gildea, 217–244. Amsterdam: John Benjamins.

Haude, Katharina. 2019a. "Grammatical Relations in Movima: Alignment Beyond Semantic Roles." In *Argument Selectors: New Perspectives on Grammatical Relations*, ed. by Alena Witzlack-Makarevich, and Balthasar Bickel, 213–256. Amsterdam: John Benjamins.

Haude, Katharina. 2019b. "From Predication to Reference: On 'Verbal DPs' in Movima." In *Individuation et référence nominale à travers les langues*, ed. by Christine Bonnot, Outi Duvallon, and Hélène de Penanros, 53–77. Paris: Editions Lambert-Lucas.

Haude, Katharina. 2021. "Clefting and Nominal Predication: Two Focus-Marking Constructions in Movima." *Faits de Langues* 52 (1): 117–138.

Haude, Katharina. 2024. "Between Symmetrical Voice and Ergativity: Inverse and Antipassive in Movima." *International Journal of American Linguistics* 90 (1): 1–36.

Haude, Katharina. 2025. "An Austronesian-Type Voice System in an Amazonian Isolate? Comparing Movima and Tagalog." In *Investigating Language Isolates: Typological and Diachronic Perspectives*, ed. by Iker Salaberri, Dorota Krajewska, Ekaitz Santazilia, and Eneko Zuloaga, 71–102. Amsterdam: John Benjamins.

Himmelmann, Nikolaus P. 1991. *The Philippine Challenge to Universal Grammar* (Arbeitspapier 15, Neue Folge). Köln: Institut für Sprachwissenschaft.

Himmelmann, Nikolaus P. 2008. "Lexical Categories and Voice in Tagalog." In *Voice and Grammatical Functions in Austronesian Languages*, ed. by Peter K. Austin, and Simon Musgrave, 247–293. Stanford: CSLI Publications.

Jelinek, Eloise, and Richard A. Demers. 1994. "Predicates and Pronominal Arguments in Straits Salish." *Language* 70 (4): 697–736.

Kaufman, Daniel. 2009. "Austronesian Nominalism and Its Consequences: A Tagalog Case Study." *Theoretical Linguistics* 35 (1): 1–49.

Kaufman, Daniel. 2025. "Why is Omniargumentivity Harder Than Omnipredicativity?" Presentation, 17th Conference on Austronesian and Papuan Languages and Linguistics, Cologne.

Kinkade, M.Dale. 1983. "Salish Evidence Against the Universality of 'Noun' and 'Verb.'" *Lingua* 60 (1): 25–39.

Launey, Michel. 1994. *Une grammaire omniprédicative: Essai sur la morphosyntaxe du nahuatl classique*. Paris: CNRS Editions.

Launey, Michel. 2004. "The Features of Omnipredicativity in Classical Nahuatl." *STUF — Language Typology and Universals* 57 (1): 49–69.

Malcher, Kurt. 2021. Clefts. A Cross-Linguistic Investigation. PhD thesis, University of Cologne. https://kups.ub.uni-koeln.de/71050/

Sasse, Hans-Jürgen. 1987. The thetic/categorical distinction revisited. *Linguistics* (25): 511–580.

CHAPTER 24

Information structure and the agent-first preference

A corpus-based case study of a free word order language

Eva Schultze-Berndt
The University of Manchester

This paper investigates correlations between information structure and constituent order patterns in transitive clauses extracted from recorded texts in Jaminjung-Ngaliwurru, an Australian language of the Mirndi family without a dominant constituent order at clause level. In transitive clauses with two overt arguments, and considering only linear order, the findings support a widely reported cross-linguistic preference for orders where the agent precedes the patient. However, the prevalence of agent-first orders is an epiphenomenon of two related preferences: a high probability for topics to be agents, and a preference for a rheme-internal placement of the patient, in contiguity with the verb. The result is that agents, in the case that they are overt at all, tend to be placed in rheme-external position, usually as initial topics.

Keywords: constituent order, information structure, sentence topic, agent-first preference, Australian languages

1. Introduction

Clause-level constituent order has been a highly prominent topic in linguistic typology since Greenberg's (1966) seminal work. Research based on data on dominant constituent orders — e.g. on figures in the World Atlas of Language Structures (Dryer 2013a, 2013b, 2013c) — has resulted in far-reaching proposals not only regarding the correlation of constituent order and other grammatical features such as case-marking (e.g. Sinnemäki 2010) but also regarding tendencies of language change (e.g. Riesberg et al. 2019) and even language evolution (e.g. Gell-Mann & Ruhlen 2011; Bickel et al. 2015).

https://doi.org/10.1075/slcs.240.24sch
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

A much-discussed claim is a universal preference for linear orders where the agent (transitive "subject") precedes the patient ("object") in transitive clauses, or agent-first preference for short, as already formulated in Greenberg's Universal 1 (1966:77). Evidence cited in support of this preference comes from typological datasets such as Dryer (2013a) (where it corresponds to the dominant order in 96.6% of languages classified as having a dominant order) as well as from studies of languages with a specific typological profile such as symmetrical voice languages[1] (Riesberg & Primus 2015; Riesberg et al. 2019). Numerous linguistic production experiments for a variety of languages (see Section 7 for further discussion and references) and even experiments on non-linguistic ordering preferences, e.g. in pantomime representations of scenes (e.g. Goldin-Meadow et al. 2008; Hall et al. 2013), have likewise supported the agent-first preference. Neurolinguistic perception studies reveal an agent-first bias in the sense that there are observable neurophysiological effects of increased processing effort whenever an ambiguous initial argument has to be disambiguated to a non-agent. The agent-first preference is claimed to hold even for ergative languages such as Hindi or Basque (Bickel et al. 2015; Huber et al. 2024), for verb-initial languages like Kaqchikel (Koizumi 2023: Ch. 9) and — in the case of the Australian language Murrinhpatha — for a language for which no dominant clausal constituent order can be established (Nordlinger et al. 2022).

Although most of the research on the agent-first preference concentrates on the first argument in a clause, I will take as basis for the discussion the more general formulation in Riesberg et al. (2019:525):

> Whenever word order in transitive constructions allows for different options, the pool of alternative options will be strongly biased toward an 'agent-first' order.

This predicts that agents have a tendency to precede patients wherever their position in the clause, within a given language as well as across languages or varieties.

Meanwhile, it is well known that transitive clauses with two overt arguments are actually very rare in informal spoken discourse. This dispreference for more than one lexical argument per clause is captured by Du Bois' (1987, 2003) Quantity Constraint. Moreover, in many, perhaps all languages, flexibility in clause-level constituent order is exploited for discourse-pragmatic reasons: when planning utterances, speakers do more than just decide on a linear order; they also select an information structure pattern which may be reflected in constituent order and/or prosodic encoding, syntactic constructions, and morphological marking. This is true even for "configurational" languages with relatively fixed constituent order, but

1. In these languages, the notions of agent and subject clearly diverge in Undergoer voice constructions, and the evidence points towards an agent-first rather than subject-first preference.

more pronounced for languages with flexible order (leading to the label "discourse-configurational").

This observed language-internal variability has led to calls for a gradient approach to constituent order based on text frequencies and contexts of use (e.g. Wälchli 2009; Haig et al. 2011; Schnell & Schiborr 2022; Levshina 2019; Levshina et al. 2023) — in other words, based on primary data rather than structural data (Himmelmann 2012). The inclusion even of lesser-studied languages in such an undertaking has been made possible by documentary linguistics and its focus on audio-visual corpora of languages, to which Nikolaus Himmelmann has contributed enormously, both in theory (e.g. Himmelmann 1998, 2006a, 2006b, 2012, 2016) and in practice.

The present paper is a case study contributing to discourse-based typology in that it explores the frequencies of clause-level constituent order patterns in a language without a dominant clause-level constituent order, Jaminjung-Ngaliwurru (Mirndi family, Australia). Moreover, it goes one step further in aiming to establish correlations of different orders with specific information structure patterns in a small corpus annotated for information structure functions such as topic and focus, based on contextual and prosodic criteria. The findings highlight the importance of also including basic prosodic information in the annotation of spoken language corpora (Himmelmann 2006b; Simard & Schultze-Berndt 2011).

The focus in this study will be on the small subset of transitive clauses with two overt core arguments ("subject"/A and "object"/P) in the corpus, since these are the type of clauses used in the literature to test for the agent-first preference. As we will see, however, the very low frequency of this type of clause raises questions about the relevance of an agent-first preference as a processing preference (see further Section 7).

The paper is structured as follows: Section 2 provides an introduction to Jaminjung-Ngaliwurru and its main grammatical properties, including the criteria employed in the study for identifying transitive clauses. The criteria for identifying basic information structure categories are outlined in Section 3. Section 4 describes the corpus and the dataset of transitive clauses. Constituent orders and argument realisation patterns are discussed in Section 5 for all transitive main clauses in the corpus, whereas Section 6 examines the correlations of constituent order and information structure in the core dataset of clauses with two overt arguments. The findings are discussed in Section 7 with a focus on the agent-first bias and other claims about cross-linguistic preferences. In brief, I will argue that the agent-first preference in the attested frequency of constituent orders, at least for Jaminjung-Ngaliwurru, is an epiphenomenon of a tendency for initial topics to be agents. I will also suggest that the observed patterns have a better explanation in a preference for the contiguity of verb and object, which leads to the place-

ment of agents in rheme-external position (to the extent that they are overt at all). Section 8 concludes with a brief summary of the arguments and evidence and an outlook to further research.

All examples in this paper come from the sub-corpus used for this study and are annotated for the information structure categories further detailed in Section 3.

2. Language background

Jaminjung and Ngaliwurru are two closely related and mutually intelligible named varieties spoken by a few elderly speakers in various settlements in the Victoria River area in northern Australia. Together, they form the western branch of the discontinuous Mirndi language family (Chadwick 1997; Green & Nordlinger 2004; Harvey 2008). The language of daily interaction in the area is Kriol, an English-lexified creole language.

The differences between Jaminjung and Ngaliwurru are mostly of a lexical nature and only marginally phonological and morphological; as far as constituent order and information structure are concerned, the two varieties do not differ. The main grammatical characteristics that will be relevant for the discussion of constituent order are described in Section 2.1. Section 2.2 outlines the criteria for the identification of transitive clauses in the corpus.

2.1 Grammatical overview

Jaminjung-Ngaliwurru is double-marking in the sense of Nichols (1986): overt noun phrases in core argument function are case-marked following ergative-absolutive alignment, albeit with variable ("optional") ergative marking of agents (Schultze-Berndt 2017a; Schultze-Berndt & Meakins 2017). Absolutive case is unmarked. Core arguments, in addition, are obligatorily indexed by pronominal prefixes representing the single argument with intransitive verbs, and agent (A) and patient/recipient (P) with transitive and ditransitive verbs.[2]

Examples (1) to (3) provide an initial illustration of double argument marking as well as of variable constituent order. They also illustrate additional enclitic pronominals (*=yirrag* and *=yirri)* which index oblique arguments (beneficiaries/locations and maleficiaries). The examples are from the same text (and a single speaker), which reports on the living conditions in the speaker's small commu-

2. In this paper, I use the abbreviations A and P interchangeably with the labels *agent* and *patient* to indicate macro-roles with core argument status.

nity when it was first set up. The orders illustrated are AVP (1), APV-A (2),[3] and PAV (3).

(1) [[*Yulngu=**ni***]$_{F}$ ***gana**-nthama=yirrag* *<u>bulayi</u> gujugujugu*]$_{RH}$
name=ERG 3SG>3SG-bring.IPFV=1PL.EXCL.OBL tent RDP:big
'The Yolngu organisation brought us big (sheets for) fly tents'
(DB, ES96_A10_02.062)

(2) [*Girrang* [*jungulug=**di***]$_{F}$ *<u>haus</u>* ***gana**-ma-na=yirri*]$_{RH}$,
yet one=ERG house 3SG>3SG-have-IPFV=1PL.EXCL
[*bojwan=**ni**=wung*]$_{AFT}$
boss=ERG=RESTR
'As yet (just) one (person) had a house "on us", just the boss'
(DB, ES96_A10_02.103)

(3) [*<u>Haus</u>=malang*]$_{TOP}$ [[*bojwan=**ni**=wung*]$_{F}$ ***gana**-ma-na=yirri*]$_{RH}$
house=GIVEN boss=ERG=RESTR 3SG>3SG-have-IPFV=1PL.EXCL
'A house only the boss had "on us"' (DB, ES96_A10_02.094)

With ditransitive verbs of the double-object type (Schultze-Berndt 2010), both theme and recipient appear in the absolutive, but it is the recipient which is treated as P in the verbal indexing system. Consequently, the recipient will be treated here as the primary object in the sense of Dryer (1986), and the theme as the secondary object (P2), in the expected case where the recipient is animate and the theme is inanimate (no exceptions occurred in the core dataset). Reflexive and reciprocal constructions are formed by a derivational suffix which renders the verb morphologically and syntactically intransitive (Schultze-Berndt 2023); as a result, no reflexives are included in the datasets analysed in this study. Jaminjung-Ngaliwurru has no voice (passive or anti-passive) constructions, and no noun incorporation.

Arguments do not have to be represented by nominal expressions if their referent can be retrieved from the discourse context or is a speech act participant, as illustrated in (6b) (see further Section 5). In fact, clauses frequently consist only of a (simple or complex) verb. Correspondingly, free pronouns are only used in emphatic contexts as illustrated in (4) below (cf. Baker & Mushin 2008: 11), and are rare in discourse (see further Section 5).

Following Schultze-Berndt & Simard (2012), I assume the existence of a phrasal category noun phrase for Jaminjung-Ngaliwurru, on the grounds of prosodic criteria and the possibility of phrase-level case marking. Exponence of a

3. As will be explained in Section 4.2, this clause was excluded from the core dataset because the A argument is represented twice.

single referent by multiple NPs is relatively frequent in discourse, e.g. when one of them elaborates on the other, in the form of an afterthought, as in (2) (see further Sections 3.3 and 4.2). Again following Schultze-Berndt & Simard (2012), such instances are distinguished from discontinuous NPs in the narrow sense, which are extremely rare and restricted to the information structure configurations of thetic ("all-new") clauses and focused attributes. Discontinuous NPs did not occur in the core data set of transitive clauses used in this study.

Similarly to a number of other languages in the area, Jaminjung-Ngaliwurru has two "verbal", i.e. primarily predicative, parts of speech (Schultze-Berndt 2000, 2003, 2015, 2017b). The first is a closed class of inflecting verbs, semantically generic lexical items which obligatorily inflect (by affixation and/or stem suppletion) for person/number and for tense/aspect and modality, as already illustrated in (1) to (3). The second verbal part of speech is an open class of semantically specific lexemes lacking verbal inflections, termed *uninflecting verbs* or *coverbs*, e.g. *burlug* 'drink' in (4), *ning* 'break' in (7), and *jag* '(move) down' in (11). These combine with inflecting verbs to form complex predicates which are usually contiguous (albeit with variable order). While arguments or adjuncts can occasionally intervene between the two verbal components, this did not occur in the core dataset for this study. The abbreviation V in this paper thus stands for either a simple inflecting verb or a complex predicate.

2.2 Defining transitive clauses

In order to identify the frequencies of different constituent orders in transitive and ditransitive clauses in Jaminjung-Ngaliwurru, clear criteria are required. For the purposes of this study, a clause was classified as transitive if the simple or complex verb potentially (i.e. as evidenced in the wider corpus and lexicographic description) co-occurs with two core NPs, one (P) in the unmarked absolutive case and the other (A) variably absolutive or ergative-marked. In order to qualify as a ditransitive construction, both P (primary object) and P2 (secondary object), if overt, had to occur in the absolutive (Schultze-Berndt 2010).

The definition is based on the actual or potential overt occurrence of two (or three) syntactic arguments and on core case marking, rather than on indexing on the verb, because there are a number of regular mismatches between morphological and syntactic transitivity (see Schultze-Berndt (2015) for details). For example, some morphologically transitive inflecting verbs, such as the speech framing verb *-Junggu* 'say/do', have conventionalised readings which do not permit an NP as P argument and have a non-referential (invariable 3rd person singular) object index.

Conversely, a transitive clause need not involve a morphologically transitive verb. For example, the morphologically intransitive verb *-irna* 'burn' allows for two core syntactic arguments. Also, Jaminjung-Ngaliwurru has a bivalent, bi-absolutive pseudo-progressive construction (Schultze-Berndt 2012) formed with a morphologically intransitive verb, as shown in (4). If two absolutive arguments were present with this construction, they were identified as A (*burri* '3PL' in 4) and P (*nalija* 'tea' in 4), respectively, on the basis of their semantic roles.

(4) [[*BURRI=biyang*]$_{F}$ *burlug=mayan burru-w-iyaj*]$_{RH}$ \\ [*nalija*]$_{A\text{-}T}$
3PL=SEQ drink=ITER 3PL-POT-be tea
'(When the three will come back), THEY will drink tea then'
(JosJ, ES15_A02_02.037b)

A case of apparent mismatch between the number of overt core NPs and semantic (and morphological) valency is presented by external possession ("possessor raising") constructions. In Jaminjung-Ngaliwurru, like in many languages in the area (Ponsonnet 2023: 248–251), both an affected part — usually a body part — and its possessor can appear as unmarked (absolutive) noun phrases in a single clause in configurations of inalienable possession of the part-whole type. The possessor is treated as the intransitive subject or object[4] in terms of indexing on the verb, as shown in (5).

(5) [*Gurrany* ***gan****-ngangu* *juwud*]$_{RH}$ \\ [*warnanggal=ni*]$_{AFT}$
NEG **3SG>1SG**-get/handle.PST eye doctor=ERG
'He didn't treat my eyes, the doctor' (lit. 'He did not treat me (with respect to) eyes') (MM, CS15_A014_22.031–032)

The syntactic analysis of such constructions is a matter of debate in the literature; see Ponsonnet (2023: 246) for an overview. I consider the possessed part expression not as a syntactic argument but as an additional noun phrase licensed by the possessive relationship. Therefore, I did not count phrases such as *juwud* 'eye' in (5) as arguments when identifying transitive clauses with two core arguments (their inclusion would have resulted in no more than 8 additional clauses in the core dataset).

4. Overtly represented body parts possessed by an A were considered as instruments, i.e. obliques, for the purposes of this study (ergative and instrumental case are identical in Jaminjung-Ngaliwurru).

3. Information structure in Jaminjung-Ngaliwurru

In this section, the basic information structure categories used in this study will be defined, with a focus on their language-specific correlates. This builds on research on the prosody and information structure of Jaminjung-Ngaliwurru by Simard, as well as collaborative work (Simard 2010, 2013a, 2013b, 2014; Simard & Schultze-Berndt 2011; Schultze-Berndt & Simard 2012). In line with many other accounts (e.g. Lambrecht 1994), information structure will be distinguished from referential status in discourse, as captured by categories such as givenness (previous mention), accessibility, specificity, or definiteness. While there may be correlations e.g. between topic position and high accessibility, and discourse-new referents and focal status, these are not absolute. Referential status was not systematically investigated in this study, though it may be commented on for individual examples where relevant.

3.1 Rheme (comment) vs. focus

Each utterance minimally consists of a rheme or comment (Hockett 1958: 201). This is the primary unit of information which bears illocutionary force, and it is therefore the core and only obligatory component of any independent utterance. The term *rheme*, while less established than *comment*, is used here because it does not evoke the necessity of a complementing topic/theme. Example (6) provides an illustration: while the rheme may be preceded by an overt initial topic (6a) (see Section 3.2) or followed by an afterthought (6a) or an anti-topic (6c) (see Section 3.3), in naturalistic discourse rheme-only clauses frequently occur (6b), usually — as in this case — as a comment on an implicit topic.

(6) a. [*Galbun=ni*]$_{\text{TOP}}$ // [*juru gan-arran-y yinthu*]$_{\text{RH}}$ // [*dimarlan=ni*
whistling.kite=ERG nest 3SG>3SG-put-PST PROX white.gum=LOC
garndi=ni]$_{\text{AFT}}$ \\
tree=LOC
'The whistling kite, she made a nest here, in the white gum tree'

b. [*Juru gan-arra::-ny*]$_{\text{RH}}$ // [*marndaj*]$_{\text{RH}}$ \\
nest 3SG>3SG-put-PST all.right
'She made a nest, all right'

c. [*Thamang bagurr ga-rdba-ny*]$_{\text{RH}}$ \\ [*gujarding=malang*]$_{\text{A-T}}$ \\
finished be.inside 3SG-fall-PST mother=GIVEN
'Afterwards she got inside, the mother (kite)' (DR, CS15_A013_25.002-6)

The notion of rheme is distinguished from the notion of *focus* as follows: a focus expression fills the open variable in an implicit or explicit question under discus-

sion (QuD) and thus evokes the presence of alternatives that are relevant for the interpretation of the expression at hand (Dik 1997: 328; Krifka & Musan 2012: 7). It can be co-extensive with the rheme (broad focus) or with a constituent within the rheme (narrow focus). Broad focus of the predicate focus type corresponds to a QuD such as 'what is X doing' or 'what is happening to X', where X is the referent of an overt topic, as e.g. in (3), (6a), and (10), or of an implicit topic, as in (6b). If the QuD concerns the development of events in their entirety ('What happens/happened?'), and all arguments are included in the rheme, the broad focus is of the thetic (or sentence focus) type (Kuroda 1972; Lambrecht 2000; Sasse 1987, 2006), as in (7). Clear cases of narrow focus — where the QuD concerns the identity of one of the participants in an event — are noun phrases that correspond to a constituent interrogative in a question-answer pair, as in the staged dialogue in (8), or that clearly contrast with alternatives, as in (9). In this case, the rheme is subdivided into focus and background (Krifka & Musan 2012: 28).

(7) [*Wajgany ning gani-wa wirib=ni*]$_{\text{RH}}$ \\
honey/beehive break.off 3SG>3SG-bite.PST dog=ERG
'The dog bit off the beehive!' (IP, ES97_A03_01.178)

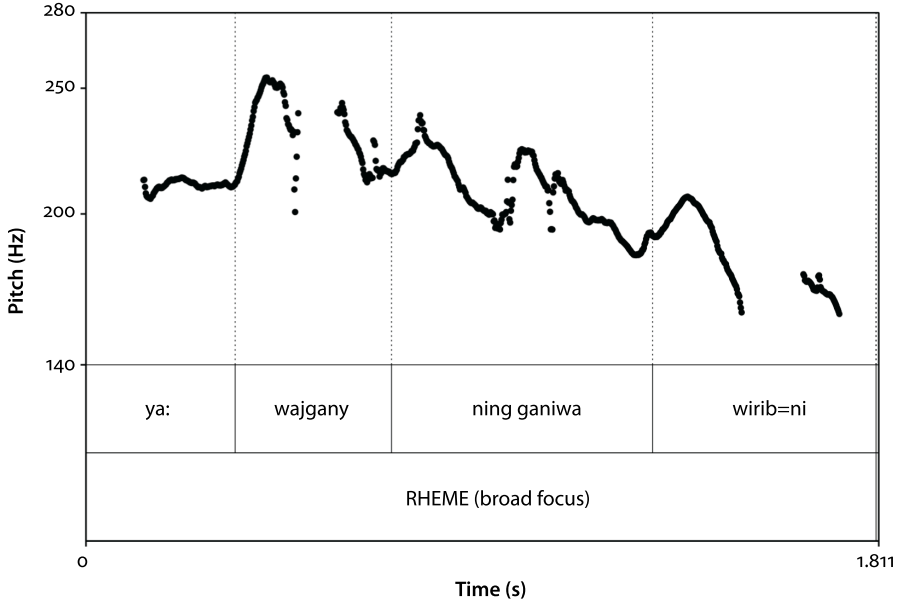


Figure 1. Pitch contour for Example (7)

(8) a. [[*NANGgarni*]$_{F}$ *gan-uga=rrgu* *ngarrgina dubuluj?*]$_{RH}$
who:ERG 3SG>3SG-take.PST=1SG.OBL 1SG:POSS bag
'Who took my bag from me?'

b. [*aa majani* [*jarlig=burlu=ni*]$_{F}$ *burr-uga*]$_{RH}$ \\
ah maybe child=COLL=ERG 3PL>3SG-take.PST
'Ah, maybe the children took it!'

c. *yawayi, yawayi,* [*burr-uga* [*jarlig=burlu=ni*]$_{F}$]$_{RH}$ \\
yes yes 3PL>3SG-take.PST child=COLL=ERG
'yes, yes, the children took it' (JM+ERa; ES08_A20_03.080–86)

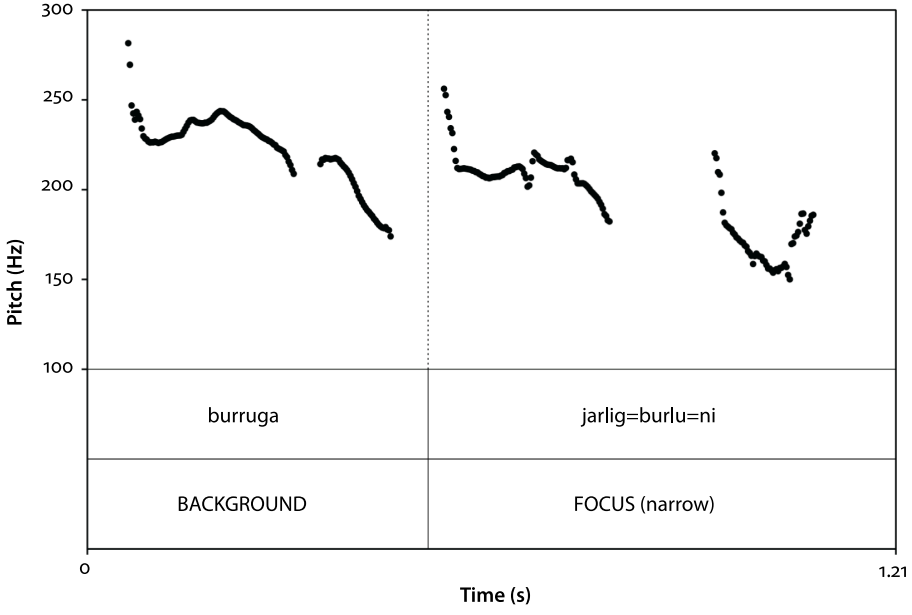


Figure 2. Pitch contour for Example (8c)

(9) [[*Burrgal*]$_{F}$ *ngayug nga-rra-m*]$_{RH}$ \\
black.plum 1SG 1SG>3SG-put-PRS
'*Burrgal* (black plum) I call it.' (DB, ES94_A08_01.183)

In Jaminjung-Ngaliwurru, the main and most salient prosodic correlate of information structure is a steeply falling contour, the onset of which is aligned with the first lexical prosodic word in the scope of focus (Simard 2010: 221,296).[5] If the entire rheme is focal, the falling contour is thus aligned with the beginning of the

5. Note that the final boundary of a focal contour can show a continuing fall, but also a rising or level intonation. These final contours encode separation vs. continuation, and do not participate in the marking of information structure (Simard 2013a, 2010: 325–326).

rheme (excluding unstressed particles). This can be seen in Figure 1 (Example (7), one of the few examples of sentence focus in the dataset)[6] and also in Figure 3 (Example (10)) and Figure 4 (Example (12)). For postverbal narrow focus as in (8c), a pitch reset and falling contour will mark the focal constituent, as can be seen in Figure 2.

As Examples (8) and (9) show, there is no fixed position for constituents in narrow focus in Jaminjung-Ngaliwurru. While postverbal narrow focus as in (8c) can be clearly identified prosodically, the identification has to rest on contextual criteria for rheme-initial narrow focus. There are also no necessary morphosyntactic or prosodic correlates of sentence focus to distinguish it from predicate focus, though see Simard (2010: 225–234), Schultze-Berndt & Simard (2012) and Schultze-Berndt (2022) for further discussion.

3.2 Initial topic

The identification of topics in this study is based on a generally accepted definition of a (sentence) topic as an overt constituent denoting what the sentence is about (Lambrecht 1994; Krifka & Musan 2012: 27). Jacobs (2001) makes the notion of aboutness more precise in two ways: firstly, building on Hockett's (1958: 201) original definition that the speaker "announces" a topic and then says something about it, Jacobs describes aboutness by the metaphor of providing an explicit "address" under which the subsequent information is to be stored (2001: 650–655). This is compatible with the proposal that topics have their own illocutionary force (Portner & Yabushita 1998; Krifka 2001). Secondly, Jacobs (2001: 645–647) proposes the additional criterion of informational separation of the topic constituent from the comment (rheme). This conveniently translates into a formal criterion of positioning and prosodic phrasing: it entails that topical constituents occur at the edge of clauses — either as initial topics (preceding the rheme) or anti-topics (following the rheme) — and are separated from the rheme by a prosodic boundary.

In Jaminjung-Ngaliwurru, initial topics are followed by a prosodic break, usually consisting of a following major pitch reset associated with a focal pitch contour coinciding with the beginning of the rheme (see Section 3.1), and optionally a pause. The pitch reset (on the complex predicate) is illustrated for Example (10) in Figure 3. The prosodic contour of the topic constituent itself can be variable; while it often ends in a rise (indicating continuation), this is not necessarily the case (Simard 2010: 261,270).

---

6. While the dog in (7) is not discourse-new, the speaker is commenting on a new scene in the Frog Story which presents a surprising development.

(10) [*Jalig=di=malang*]$_{\text{TOP}}$ [*wirr gan-ardgiya-ny=nu*]$_{\text{RH}}$ \\ [*mali*]$_{\text{AFT}}$ \\
child=ERG=GIVEN move.out 3SG>3SG-throw-PST=3SG.OBL thing
'The child shook it out, the thing (i.e. boot)'
(Frog Story; CP, ES96_A18_02.018)

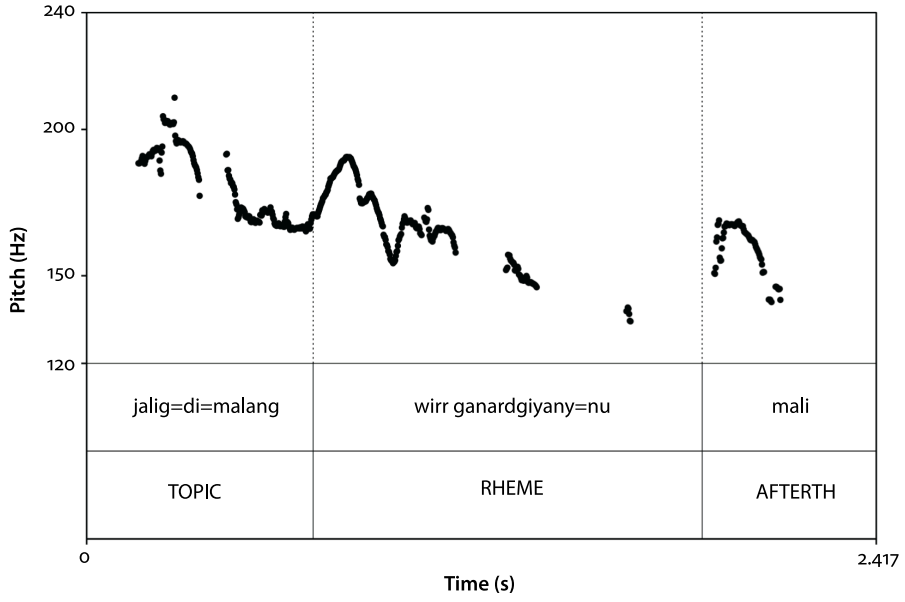


Figure 3. Pitch contour for Example (10)

Topics as defined here are not necessarily to be equated with "old" (i.e. discourse-given) information. Discourse-old referents tend to be omitted in Jaminjung-Ngaliwurru, as in many other languages (see also Section 5). Overt initial topics can introduce new referents, as in Example (6a) (which is the first utterance of a text), but also signal a topic shift between previously established referents, as in (10).

3.3 Anti-topic and afterthought

Rheme-external constituents — in the sense that they are separated from the rheme by a clearly audible prosodic boundary, signalled by a boundary contour and usually a pause — can also follow the rheme. The term *anti-topic* has been coined (Chafe 1976; Lambrecht 1981, 1994) for post-rhematic units with the function of signalling the reactivation of a discourse referent rather than the shift to a new discourse topic (Ziv 1994; Lambrecht 1994: 203, 2001; Gregory & Michaelis 2001; Averintseva-Klisch 2008). For example, in (11), from a Frog Story retelling,

the dog is mentioned in order to reactivate it as a referent, even though it has been a protagonist from the beginning of the narrative, and it has been mentioned before several times. This example also demonstrates that both core arguments can be expressed in rheme-external positions, in this case topic (P) and anti-topic (A).

(11) [*janju=malang mulanggirrng*]$_{\text{TOP}}$ [*jag* <u>*na*</u> *gan-angu*]$_{\text{RH}}$ \\
DEM=GIVEN dangerous down now 3SG>3SG-get/handle.PST
[*wirib=ni=malang*]$_{\text{A-T}}$ \\
dog=ERG=GIVEN
'those dangerous ones (i.e. bees), it got them down, the dog'
(CP, ES96_A18_02.069)

The rhematic unit can also be expanded by an *afterthought* (e.g. Aijmer 1989; Ziv 1994; Auer 1996; Averintseva-Klisch 2008): a non-topical constituent which follows a prosodic boundary and serves to complete (or elaborate on) the answer to a QuD (question under discussion, see Section 3.1). Afterthoughts should therefore be considered focal, and correspondingly, in Jaminjung-Ngaliwurru — as well as cross-linguistically (Grosz & Ziv 1997) — they exhibit a pronounced pitch reset and a focal falling contour which distinguishes them from Anti-Topics (Simard 2010: 276–281,314–316, 2014).

Example (12) illustrates an A in the function of afterthought, in VPA order. The corresponding pitch contour, illustrating the pitch reset and falling contour on the afterthought, is shown in Figure 4. Another example can be seen in (10) (Figure 3).

(12) [*Wirriny-wirriny gan-arra-m ngayin*]$_{\text{RH}}$ \\ [*Nangari=ni*]$_{\text{AFT}}$ \\
RDP-turn 3SG>3SG-put-PRS meat <subsection>=ERG
'She turns over the pieces of meat, Nangari does.' (MM, CS15_A014_16.002)

As we have seen, afterthoughts or anti-topics can be the sole exponent of an argument in a clause, as in (6c), (11) and (12), or they can elaborate on an argument present in the rheme, as in (2).[7] Afterthoughts can also be stacked, as in (13), from the same picture-based elicitation as (8).

(13) [*burr-anja*]$_{\text{RH}}$ \\ [*jarlig=ni*]$_{\text{AFT}}$ \\ [*jirrama* <u>*beg*</u>]$_{\text{AFT}}$ \\
3PL>3SG-take.PRS child=ERG two bag
'They are taking them, the children, two bags.' (JM, ES08_A20_03.097)

7. As will be explained in Section 4.2, clauses with dual representation of an argument, such as (2), were excluded from the core dataset.

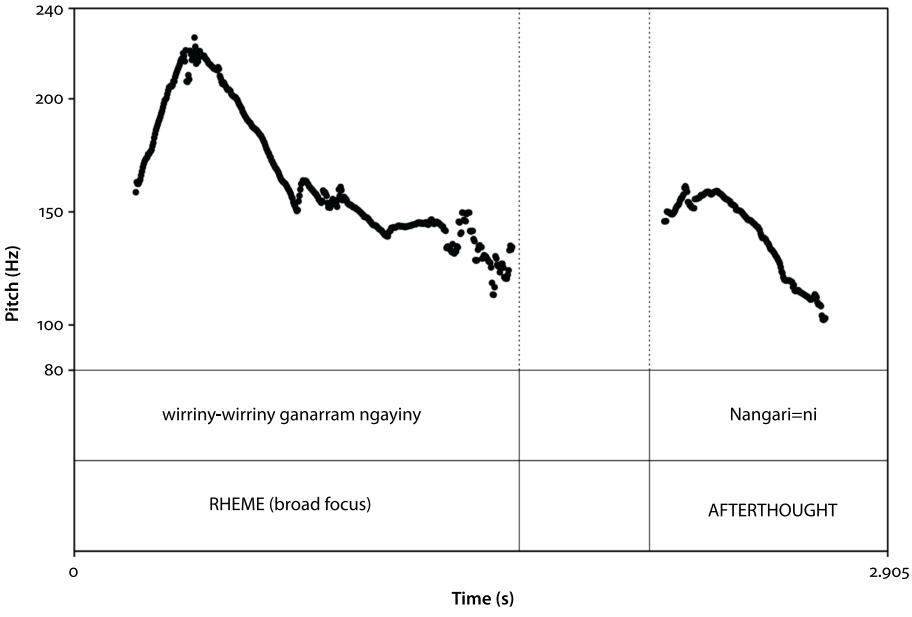


Figure 4. Pitch contour for Example (12)

3.4 Information structure and "default" constituent order

It could be claimed that clauses with rheme-external positions as discussed in 3.2 and 3.3 — i.e. with (initial) topics, anti-topics, or afterthoughts — constitute marked clause types that are derived from clauses with a basic or default constituent order. Quite apart from the issue that no such default order can be established (see below Section 6), I argue that no marking is involved beyond the prosodic partitioning of such clauses. Specifically, the rheme-external positions are not dislocated in the sense of requiring a resumptive pronoun within the rheme (other than person marking on the verb which is obligatory in any case). In an exclusively spoken language, partitioning information into shorter chunks can be considered as the norm, and the presentation of two core arguments within a single prosodic contour (as in Examples (1), (7), (8a), and (9)) as the exception; this is certainly the case for Jaminjung-Ngaliwurru (see Sections 5 and 6.2 for frequency data). In the remainder of this paper, I therefore consider all information structural patterns discussed in this section as constructions with equal, basic status — though in Section 6.2, I will consider the relevance of the informational separation of arguments as topics, afterthoughts and anti-topics for constituent order preferences.

4. Data and coding

4.1 Corpus

The current study is based on a corpus of 37 texts of varying lengths, representing 13 speakers (1 male, 12 female), which form a subset of the entire documentary corpus of Jaminjung-Ngaliwurru (Schultze-Berndt et al. 2017) and were recorded between 1994 and 2016.[8] Out of the 13 speakers, 7 reported speaking Ngaliwurru and 6 reported speaking Jaminjung. The subcorpus consists of personal and mythological narratives, procedural texts, elicitations by means of visual stimuli such as picture books, photos, and videos, and (to a lesser extent, interspersed with the other text types) elicitations of a lexicographical or syntactic nature, using English/Kriol as a metalanguage.[9] To a large extent, these are thus recordings of staged speech events in the sense of Himmelmann (1998). The recordings also include a number of short spontaneous or staged conversations, as well as reported conversations within narratives; however, conversations are not prominently represented in the data, because the language was rarely used as a medium of everyday interaction, and due to ethical considerations. Excluding contributions by the researchers and Kriol or English utterances, the total number of ELAN segments[10] in the subcorpus is 2317 (1437 by Ngaliwurru speakers and 880 by Jaminjung speakers). An overview of the texts is provided together with the archived dataset (Schultze-Berndt 2025a). Example references indicate speaker initials and a unique identifier for the text and the numbered ELAN annotation segment. The texts are audio-aligned in ELAN and fully annotated with a text tier encoding minimal prosodic features and pauses in a conversation-analysis style transcription, morphological breakdown, interlinear glosses and free English translation, and where applicable, notes on the context and speakers' commentaries. They were further coded in ELAN using the GRAID (Grammatical Relations and Animacy in Discourse) annotation conventions (Haig & Schnell

8. The majority of these texts were recorded, transcribed and annotated by myself. A few of the recordings were generously made available by Candide Simard and Mark Harvey (the initials of the person responsible for the recording correspond to the first two letters of the file name); transcriptions and annotations in these cases were undertaken or checked by myself.

9. Elicitation stimuli include the Frog Story picture book (Mayer 1969), retellings of EDED video stimuli created by Patrick Caudal and Robert Mailhammer (see Caudal & Mailhammer 2022), and picture books created by Candide Simard and by Carmel O'Shannessy (O'Shannessy 2004).

10. Annotation segments in ELAN usually, but not fully consistently, correspond to an intonation unit and thus differ from the GRAID annotation units. For issues of segmenting spoken language, see Himmelmann (2006c).

2014). The coding was undertaken by myself between 2017 and 2020, and checked again for accuracy in 2025. Multiple coders with sufficient familiarity with the language were, unfortunately, not available.

Following the GRAID conventions, multiple intonation units were coded as a single clause if they showed coherence both in terms of argument structure and prosodic patterns (i.e. if they exhibited one of the patterns spanning multiple intonation units discussed in Section 3); as already indicated in Section 3.5, this means that constituents were coded as overt arguments even if separated from the rheme by a prosodic boundary. Conversely, occasionally a single intonation unit spans multiple grammatical clauses, which were then coded separately.

The GRAID conventions distinguish S, A, P and P2 (secondary object) argument roles and in addition identify human 1st, 2nd and 3rd person as opposed to non-human 3rd person arguments. These codes are applied to lexical and non-lexical overt arguments as well as omitted core arguments. The coding of argument roles followed the language-specific criteria for transitivity outlined in Section 2.2. Mythical (dreamtime) animal characters were coded as human due to their human-like behaviour. Borrowings and code-switched arguments were included as long as the main predicate was in Jaminjung-Ngaliwurru.

4.2 The datasets

Out of the total subcorpus, all finite transitive and ditransitive main clauses — i.e. those containing an inflecting verb (see Section 2.1) — were extracted based on the GRAID codes and transferred to a spreadsheet. Clauses with an imperative predicate were excluded since their inclusion would inflate the number of unexpressed As. The total number of clauses after exclusions is 808 (756 transitive, 52 ditransitive). This is the dataset underlying the frequency figures in Table 1 and the discussion of argument representation vs. omission in Section 5.

Out of this set of all transitive and ditransitive clauses, those with both overt A and P were extracted based on the GRAID annotations ($n=97$). In order to facilitate representation and discussion, the dataset was further limited to cases of exactly two core arguments, thus excluding the rare cases of (a) ditransitive clauses with three overt arguments ($n=2$) and (b) the dual representation of arguments by more than one NP ($n=13$), for example an overt P represented both within the rheme and as an afterthought, as in (2). The resulting core dataset thus consists of 82 clauses.[11]

---

11. Note that these exclusions do not skew the overall results presented in Section 6: out of the 15 excluded clauses, 8 have mixed orders due to dual representation of a role (i.e. both A>P and P>A), 5 show A>P order, and 2 P>A order.

Within the GRAID layer of coding, basic information structure categories were superimposed on the coding of role and animacy for this core dataset, following guidelines developed by myself in collaboration with Stefan Schnell and Ruth Singer (Schultze-Berndt et al. 2025 [2020]). Coding identifies the major information structure categories discussed in Section 3, i.e. (narrow) focus (@F), afterthought (@FA), initial topic (@T) and right-edge or anti-topic (@R), leaving broad focus unmarked. As already indicated in Section 3, information structure is treated here as distinct from referential accessibility, which was not coded. The patterns found in the core dataset will be discussed in Section 6. The coded datasets (Schultze-Berndt 2025a) and notes on language-specific decisions regarding the GRAID annotation (Schultze-Berndt 2025b) are accessible and archived with The University of Manchester.

5. Argument realisation patterns in transitive clauses

In this section, I present an overview of the broader trends of argument expression in the full dataset of transitive and ditransitive clauses (see Section 4.2), before moving on to the discussion of constituent orders in the core dataset with two overt arguments in Section 6.

Table 1 presents the text counts and percentages for lexical, non-lexical (including pronouns, demonstratives, content interrogatives, and *janyung* 'other') and zero expression of A (transitive subject), P (primary object) and P2 (secondary object). Figures for human and non-human referents are shown separately.

As the figures in Table 1 show, overt A arguments occur in just 23.8% of all 808 transitive clauses in the dataset; the proportion of lexical As is even lower (18.1%). Jaminjung-Ngaliwurru thus confirms the well-established cross-linguistic tendency for agents to be omitted or to receive minimal, pronominal representation where grammatically required in a language ("Non-lexical A constraint", Du Bois 2003: 35–38).[12]

However, the findings also provide striking support for the claim that the non-lexical A constraint is an epiphenomenon arising from the propensity for A referents to be human, and for human referents to be discourse topics and therefore omitted (Everett 2009; Haig & Schnell 2016: 610–613). As Table 1 shows, when considering human and non-human referents separately, the relative proportions of expressed vs. omitted arguments are similar for A and P (and for P2, a role mainly found with non-human referents in the data); a chi-square test shows

12. As already indicated in Section 2.1, in Jaminjung-Ngaliwurru arguments are obligatorily represented by verbal prefixes, but free pronouns are only used for emphasis.

Table 1. Argument realisation patterns for A, P, and P2 in (di-)transitive clauses, excluding imperatives ($n = 808$)

		Overt				Omitted		Total	
		Lexical		Non-lexical					
		n	%	n	%	n	%	n	%
A	+human	91	13.2	40	5.8	558	81.0	689	100
	–human	55	46.2	6	5.0	58	48.7	119	100
	Total	**146**	**18.1**	**46**	**5.7**	**616**	**76.2**	**808**	**100**
P	+human	39	17.4	2	0.9	183	81.7	224	100
	–human	283	48.5	35	6.0	266	45.5	584	100
	Total	**322**	**39.9**	**37**	**4.6**	**449**	**55.6**	**808**	**100**
P2	+human	0		0		0		0	100
	–human	28	53.8	4	7.7	20	38.5	52	100
	Total	**28**	**53.8**	**4**	**7.7**	**20**	**38.5**	**52**	**100**

that the difference is not significant. The overall low proportion of overt As compared with Ps is thus due to the low proportion (15% in the dataset) of non-human A referents, as opposed to 72% of Ps with non-human referents. We will return to this observation in the discussion in Section 7.

The high rate of argument omission, moreover, results in a scarcity of clauses with two overt arguments in the texts: Only 12% of all transitive and ditransitive clauses (97/808, see Section 4.2) have two or more overt arguments, and only 7.3% have two lexical arguments ($n = 59$). Calculated as a proportion of both (di-)transitive ($n = 808$) and intransitive clauses ($n = 856$) in the corpus, two or more arguments occur in only around 6% of clauses. Jaminjung-Ngaliwurru thus also supports Du Bois' (1987, 2003) Quantity Constraint (see Section 1). The low proportion of clauses with two lexical arguments found in this study are comparable to those in other languages, for example the five languages cited in Du Bois (2003: 35), as well as Cayuga (Iroquoian) (Mithun 1992: 20) and Bardi (Nyulnyulan) (Bowern 2008: 69), all ranging between 2% and 7% of clauses.

The high rate of argument omission also means that in actual discourse, the choice between agent-first and patient-first orders does not arise very frequently. This raises the question — to be further discussed in Section 7 — to what extent the agent-first preference can be a fundamental principle for speech production and processing. First, in Section 6, I will establish whether an agent-first preference applies in Jaminjung-Ngaliwurru discourse, and if so, whether it is more pronounced in particular information structure configurations.

6. The interaction of constituent order and information structure in the core dataset

In this section, the constituent orders observed in the core dataset of 82 transitive clauses with both overt A and P arguments will be investigated in more detail. An overview of the frequency of different orders is provided in Section 6.1, before examining, in Section 6.2, their correlations with the information structure patterns outlined in Section 3. The focus of the investigation will be on data relevant for claims about a universal agent-first preference, to be further discussed in Section 7.

6.1 Overview of constituent orders

Table 2 presents an overview of the constituent orders attested in the core dataset of 82 clauses with (no fewer and no more than) two overt arguments (A and P). It also indicates the proportion of clauses observing the agent-first preference as defined in Section 1, i.e. any order where A precedes P (A > P). For ease of presentation, both primary ($n = 69$) and secondary objects ($n = 13$) are represented as P (see Sections 2.1 and 4.2). Adjuncts or non-core arguments are not represented.

Table 2. Constituent order patterns in (di-)transitive clauses with overt A and overt P in the core dataset ($n = 82$)

	Order	n	%	Examples
A > P	A+V+P	31	37.8%	1, 4, 7, 8a, 10
	A+P+V	17	20.7%	6a
	V+A+P	3	3.7%	13
	Total	**51**	**62.2%**	
P > A	P+V+A	22	26.8%	7, 11
	P+A+V*	6	7.3%	3, 9
	V+P+A	3	3.7%	12
	Total	**31**	**37.8%**	
Grand Total		**82**	**100.0%**	

* 5 out of the 6 clauses with PAV order follow the pattern in Example (9): an initial P2 in narrow focus followed by a pronominal A. These were produced by the same speaker within the same text, a discussion of plant names in several neighbouring languages based on an illustrated book. This pattern is thus possibly over-represented in the dataset.

The frequencies in Table 2 confirm the characterisation of Jaminjung-Ngaliwurru as a language with no dominant clause-level constituent order in clauses with two core arguments, using Dryer's (2013d) criterion: to be identified as the dominant order, the most frequent order should be at least twice as frequent as the next frequent one. AVP order has the highest overall frequency, but only accounts for 37.8% of all clauses, and is less than twice as frequent as the next frequent one (PVA, 26.8%).

Table 2 also suggests that Jaminjung-Ngaliwurru does support the agent-first preference: orders where A precedes P make up nearly two-thirds (62.2%) of transitive clauses with two overt arguments. The agent-first preference is entirely due to the predominance of A-initial orders (58.5% of the dataset); no preference can be seen in verb-initial clauses, which are however marginal in terms of frequency (only 6 in total). Clauses with P>A order, however, still make up a surprisingly large proportion (37.8%) compared to the reported marginality of such orders in constituent order typology databases (Dryer 2013a) as well as in cross-linguistic corpus studies (Gerdes et al. 2021: 19–20).

As already indicated in Section 3, moreover, the superficial, abstract orders presented in Table 2 map onto a variety of information structure constellations, including clauses with or without an initial topic, and clauses with arguments in separate prosodic phrases as afterthoughts or anti-topics. This mapping will be investigated in more detail in Section 6.2, leading to the conclusions that the apparent agent-first preference is an epiphenomenon of the high number of agents in the position of initial topic, and that an alternative principle — a preference for the contiguity of V and P — is more in line with overall discourse preferences (see Section 7 for discussion).

6.2 Mapping linear positions onto information structure categories

Table 3 provides an initial overview of the mappings between the argument roles of A and P and the main information structure categories as defined in Section 3, separately for human and non-human referents, and abstracting away from constituent orders.

The most striking difference between A and P in their propensity for different information structure roles can be seen in the proportion of initial topics: 35.4% of As are initial topics, but only 7.3% of Ps. This difference is statistically highly significant ($p < 0.00011$). Conversely, 74.4% of Ps are in rheme-internal position, but only 47.6% of As; the difference is also significant ($p < 0.03$). No significant difference exists in the propensity for rheme-external final positions, though figures are too low to test for afterthoughts and anti-topics separately. The dataset is also too small to test for the significance of animacy in the propensity for information

Table 3. Proportions of A and P with human and non-human referents in different information structure roles in the core dataset ($n = 82$)

		A			P		
		+hum	–hum	all	+hum	–hum	all
Rhematic	n	31	8	**39**	7	54	**61**
	%	37.8%	9.8%	47.6%	8.5%	65.9%	74.4%
Initial Topic	n	23	6	**29**	3	3	**6**
	%	28.0%	7.3%	35.4%	3.7%	3.7%	7.3%
Anti-topic / Afterth.	n	7	7	**14**	1	14	**15**
	%	8.5%	8.5%	17.1%	1.2%	17.1%	18.3%
Total	**n**	**61**	**21**	**82**	**11**	**71**	**82**
	%	74.4%	25.6%	**100%**	13.4%	86.6%	**100%**

structure functions per role. Unsurprisingly though, and as in the wider dataset (see Table 1 in Section 5), a large majority (74.4%) of A referents are human, whereas the vast majority of P referents (86.6%) are non-human.

So far, the figures suggest that the preference for A-initial orders is thus an epiphenomenon of the high propensity for initial topics to have the role of A, thus contributing to the high number of A-initial orders observed in Table 2.[13] In order to further corroborate this hypothesis, Table 4 shows the frequency of different constituent orders only for clauses where both core arguments are rheme-internal, as in Examples (1), (7), (8a), and (9); these make up just 34% (28/82) of the core dataset. While this low number means that the results are not fully conclusive, the figures show a striking parity between agent-first and patient-first orders within the rheme.

The low proportion of clauses with two rheme-internal core arguments shown in Table 4 also confirms that — as suggested in Section 3.4 — splitting transitive clauses with two overt arguments into at least two prosodically detached informational chunks is the unmarked configuration in terms of frequency in Jaminjung-Ngaliwurru discourse. A discourse preference against two arguments within the rheme can be considered a variation of the Quantity Constraint (see Section 1), though the latter is formulated with reference to clauses.

Moreover, as already shown by the figures in Table 3, the placement of a core argument outside the rheme is more likely to apply to A — specifically, As function

13. The converse does not hold, i.e. it is not true that the majority of As are topics — rhematic As make up for the highest proportion of As (47.6%) (Table 3).

Table 4. Constituent order patterns in (di-)transitive clauses with rhematic overt A and P arguments ($n = 28$)

	Order	n	%
A > P	A+V+P	11	39.3%
	A+P+V	2	7.1%
	V+A+P	1	3.6%
	Total	**14**	**50.0%**
P > A	P+V+A	8	28.6%
	P+A+V	5	17.9%
	V+P+A	1	3.6%
	Total	**14**	**50.0%**
Grand total		**28**	**100.0%**

as initial topics much more frequently than Ps do and thus, conversely, Ps occur rheme-internally, in contiguity with the verb, more frequently than As. Table 5 provides further detail on the preference for the contiguity of V and P in the frequency of linear orders (see Table 2 in Section 6.1). Arguably, a contiguity in linear order does not equal contiguity in processing terms if there is a prosodic break between the two constituents, i.e. if P is an initial topic, anti-topic, or afterthought, amounting to "informational separation" (see Section 3). Table 5 therefore distinguishes non-contiguous linear order from both linear contiguity but with a prosodic break between P and V and true contiguity, i.e. adjacency within the rheme. In none of the clauses categorised as "Immediate contiguity of V and P within rheme" was any intervening constituent found between V and P. The figures thus show that two-thirds of Ps appear in contiguity with the verb not just in abstract linear order, but in information-structural terms, lending support for a cross-linguistic preference for verb-object contiguity (see further Section 7).

The analyses in this section have not accounted for any factors that influence the order of V and P within the rheme, or any other rheme-internal order. As shown in Section 3.1, core arguments in narrow focus in Jaminjung-Ngaliwurru can occupy a preverbal as well as postverbal position, contributing to the variability of constituent order. The position of arguments under broad focus (compare Examples (7) and (12)) or as part of the background (compare Examples (1) and (2)) is, likewise, variable. The discussion in the final sections focuses on the distribution of arguments across the rheme vs. rheme-external positions, and leaves for future research any factors responsible for rheme-internal variation in ordering.

Table 5. Adjacency and informational unity of V and P in the core dataset

	n	%
Non-contiguous order (PAV, VAP orders; Table 2)	9	11.0%
Initial P = topic (PVA order)	5	6.1%
P = anti-topic or afterthought in AVP or VPA* order	14	17.1%
Immediate contiguity of V and P within rheme	54	65.9%
Total	82	100.0%

* In a single example of VPA order, both P and A were afterthoughts (a pattern shown in (13) for VAP order).

7. Discussion: An agent-first bias?

The examination of the information structure patterns associated with different clausal constituent orders in Jaminjung-Ngaliwurru corpus data with two overt arguments, undertaken in Section 6, now allows us to address the question of the agent-first preference in a more nuanced light: not merely in terms of the linear order of A and P, but in terms of the association of these orders with different information structural functions.

In Jaminjung-Ngaliwurru, the observed agent-first preference in terms of the frequency of linear order in corpus data, accounting for 62.2% of clauses with two overt arguments (see Table 2), does not hold equally for all information structure configurations. It does not appear to apply to clauses where both core arguments are rheme-internal (Table 4). Rather, the higher frequency of agent-initial orders is due to the high frequency of topic-initial clauses (see Table 3), which make up 43% of all clauses in the dataset (35/82), combined with the significantly higher likelihood for initial topics to have the role of A (83% of topic-initial clauses in the dataset, 29/35) than of P (17%, 6/35). Ultimately, this is likely to be related to the high proportion of agents with human referents (see Table 3), as claimed by Everett (2009) and Haig & Schnell (2016: 610–613), though figures were too low for statistical testing of this factor.

Additional evidence for the claim that topics, by default, are interpreted as A in Jaminjung-Ngaliwurru comes from the distribution of the "optional" ergative marking (Schultze-Berndt 2017a; Schultze-Berndt & Meakins 2017). Agents in topic position (as well as speech act participant pronouns) are more likely to appear in an unmarked form rather than with ergative case. This can be interpreted as an indication that the topic position allows the hearer to predict with reasonable certainty that the corresponding referent is an agent, so that case marking becomes redundant. A similar situation is found in other fluid ergative

languages, e.g. Tima (Schneider-Blum & Hellwig 2018; Schneider-Blum 2023; see also Riesberg et al. 2019: 550).

Himmelmann & Primus (2015: 50) provide an explanation for the agent-first preference in terms of a broader preference for edge placement of agents, claiming that "the first and last elements in a domain are more salient than middle elements." There is a competing and arguably more compelling explanation for preferred edge placement of agents than their salience, however. At this point it is important to recall (see Section 1) that overt agents are cross-linguistically rare in discourse. Jaminjung-Ngaliwurru, where agents are omitted in 76.2% of all transitive or ditransitive clauses (see Table 1 and the discussion in Section 5), is no exception. In other words, the discourse-preferred configuration (in the case of transitive clauses) is for the core rhematic unit to consist either of a predicate alone, or a predicate in combination with P. Mostly, A is an implicit (discourse) topic for such clauses. The Jaminjung-Ngaliwurru data also support the further claim that this is due to the high proportion of human referents in the agent role combined with the higher likelihood of human referents to be discourse topics.

If an A is explicitly mentioned, it is often to indicate a shift of discourse topic, e.g. when there is a contrast of two protagonists (achieved by placing A in an initial topic position), the reactivation of a discourse topic (anti-topic position) or to fully answer a Question under Discussion (afterthought) (see Section 3). This accounts for the 52% of overt As in rheme-external position in the core dataset, as opposed to only 26% of Ps (see Table 3 in Section 6.2).

The above observations cast doubt on the claim that the agent-first principle is a fundamental principle of language processing. Rather, they invite an explanation for constituent order preferences which is based on the combination of two tendencies: a preference for the contiguity of V and P (verb and object), and a higher proportion of overt agents than patients to have the information structure role of topic. The first principle has indeed been repeatedly invoked in the literature to account for cross-linguistic constituent order tendencies: it has been labelled the principle of verb-object-bonding (Tomlin 1986), the phrase structure principle (Whaley 1997: 94), the verb-object constraint (Baker 2009a, 2009b), and the VP-principle (Riesberg et al. 2019: 548, citing D. Kaufman p.c.). This principle finds a much stronger support in the Jaminjung-Ngaliwurru corpus data than the agent-first preference: in terms of purely linear order, only 11% of the core dataset (the 9 clauses with VAP and PAV order) do not obey it (Table 2). In information-structural terms, Ps are not only more likely to be overt at all than As (for Jaminjung-Ngaliwurru see Table 1), but also more likely to occur within the rheme, i.e. in broad or narrow focus, and in immediate contiguity with the verb (see Table 5).

In this study, no clear conclusions can be drawn for verb-initial clauses — the type of particular interest for the Austronesian symmetrical voice languages investigated by Riesberg et al. (2019). While no agent-first preference can be established for verb-initial clauses in Jaminjung-Ngaliwurru (see Table 2), verb-initiality in general is dispreferred in this language and overall figures are very low.

The findings presented here have implications for experimental research on the agent-first preference, although it is beyond the scope of this paper to spell them out in detail. Much of this research disregards information structure: the processing of sentences out of context (in comprehension experiments) or the description of isolated visual scenes designed to elicit "complete" transitive clauses (in production experiments) is assumed to reflect pragmatically neutral, default utterances (e.g. Koizumi et al. 2014). As has been observed before (e.g. Mithun 1992: 21; Lambrecht 1994: 190; Riesberg et al. 2019: 525, fn 5), however, clauses with two overt (lexical) arguments, far from representing a pragmatically neutral preferred constituent order, can be considered a marked phenomenon in themselves, due to their very low discourse frequency. The findings for Jaminjung-Ngaliwurru indicate that in clauses with two core arguments in coherent discourse, As are very likely to be placed rheme-externally, especially as initial topics. In experimental research, it is therefore important to take into account what strategies are chosen by participants to package (or interpret) the information presented. In comprehension experiments on the agent-first preference, with few exceptions (e.g. Sauppe et al. 2023), speakers are presented with written sentences out of context — a highly artificial condition which does not mirror actual communicative situations. It has been argued (Sedarous & Namboodiripad 2020: 4) that participants impose a default prosody on such written stimuli, which moreover could differ between participants. It is not implausible that a topic-rheme prosodic pattern is a default strategy for out-of-context sentences presented in written format, which would favour agent-initial orders.

Where information structure and discourse context has been taken into account in experimental research (e.g. Kaiser & Trueswell 2004; Hung & Schumacher 2012), it has been shown that providing an appropriate discourse context for object-first clauses can override any agent-first preference. Similarly, at least for some languages, the agent-first preference could only be upheld for animates/humans (Sauppe et al. 2023; Egurtzegi et al. 2025), possibly because speakers interpreted these, but not the agents with inanimate referents, as topics.

8. Conclusions and outlook

This study contributes to the nascent field of corpus-based typology by systematically exploring the correlations between constituent orders and information structure patterns in a lesser studied language with "free" constituent order. The observed frequencies of the various constituent orders in clauses with two overt arguments in Jaminjung-Ngaliwurru confirm previously established universal tendencies, specifically, both the principle of verb-object contiguity and (less strongly) an agent-first preference. However, the agent-first preference was shown to be due to a high proportion of topic-initial clauses in the set of clauses with two overt core arguments combined with a high propensity for topics to be agents.

Due to the relatively small size of the core dataset, which reveals indicative trends of constituent placement but has insufficient statistical power, a number of questions have been left unexplored. One such question concerns the relative animacy of A and P as a potential factor; others are the factors of speaker and genre, the potential role of priming and, as mentioned in Section 6.2, the factors underlying the ordering of rheme-internal core arguments. The question of how the information structure categories investigated here are distributed in the wider corpus will, likewise, have to be left for future research. Still, the results flag the importance of taking into account information structure (as well as animacy) in designing and evaluating research on constituent order preferences in general, and on the agent-first preference in particular.

Acknowledgements

I am grateful beyond words to the Jaminjung and Ngaliwurru speakers — many of them sadly now deceased — who have taught me their language and who have over many years contributed to the documentation on which this paper is based. For financial support which has made fieldwork and documentation possible I thank the Max Planck Society and the Volkswagen Foundation (DobeS grants 82957 and 86101). My work on information structure has greatly benefitted from joint work with Candide Simard and Claudia Wegener on the latter DobeS project, and from many discussions with Dina El Zarka. The development of the information structure additions to the GRAID annotation conventions was undertaken in collaboration with Stefan Schnell and Ruth Singer and supported by the Manchester-Melbourne Humanities Consortium Fund grant "Information Packaging in corpora of lesser studied languages" (2017–2018). Sincere thanks also to the editors of this volume, Birgit Hellwig and Sonja Riesberg, as well as two anonymous reviewers, for detailed and helpful comments on a previous version of this paper. Last but not least, I would like to thank Nikolaus Himmelmann for his crucial inspiration and encouragement that made me embark on the journey of language-documentation based research.

Data availability

The annotated dataset for this paper (Schultze-Berndt 2025), in the format of a spreadsheet, is archived with the University of Manchester and freely accessible. It consists of the transcribed version of all transitive clauses found in the reference corpus with English glosses, an English free translation, an annotation following the GRAID (Grammatical Relations and Animacy in Discourse) conventions supplemented by information structure categories (Schultze-Berndt et al. 2025 [2020]), and further annotations for constituent order and information structure which are explained in the file. The same file includes metadata for the texts in the reference corpus, a list of abbreviations used in glossing, and details for the statistical tests reported in the paper. The recordings and time-aligned ELAN versions of the annotations are archived within the larger corpus (Schultze-Berndt et al. 2017) with the DobeS Archive,[14] with access provided upon request.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following abbreviations are used:

>	[A] acting on [P] (for pronominal prefixes)
\\	Final intonation unit boundary
//	Continuing intonation unit boundary
AFT	Afterthought
A-T	Anti-topic
COLL	Collective
F	Focus
GIVEN	Marker of discourse-given status
ITER	Iterative
P2	Secondary object in ditransitive predicates
POT	Potential/future verbal inflection
RDP	Reduplication
RESTR	Restrictive clitic ('just', 'right there', 'still')
RH	Rheme
SEQ	Sequential clitic ('now', 'and then').

References

Aijmer, Karin. 1989. "Themes and Tails: The Discourse Functions of Dislocated Elements." *Nordic Journal of Linguistics* 12 (2): 137–154.

Auer, Peter. 1996. "On the Prosody and Syntax of Turn-Continuations." In *Prosody in Conversation: Interactional Studies*, ed. by Elizabeth Couper-Kuhlen, and Margaret Selting, 57–100. Cambridge: Cambridge University Press.

14. https://dobes.mpi.nl/projects/jaminjung/

Averintseva-Klisch, Maria. 2008. "To the Right of the Clause: Right Dislocation vs. Afterthought." In *'Subordination' versus 'Coordination' in Sentence and Text: A Cross-Linguistic Perspective*, ed. by Cathrine Fabricius-Hansen, and Wiebke Ramm, 217–239. Amsterdam: John Benjamins.

Baker, Mark. 2009a. "Language Universals: Abstract but Not Mythological." *Behavioral and Brain Sciences* 32 (5): 448–449.

Baker, Mark. 2009b. "Formal Generative Typology." In *The Oxford Handbook of Linguistic Analysis*, ed. by Bernd Heine, and Heiko Narrog, 285–312. Oxford: Oxford University Press.

Baker, Brett, and Ilana Mushin. 2008. "Discourse and Grammar in Australian Languages." In *Discourse and Grammar in Australian Languages*, ed. by Ilana Mushin, and Brett Baker, 1–24. Amsterdam: John Benjamins.

Bickel, Balthasar, Alena Witzlack-Makarevich, Kamal K. Choudhary, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. 2015. "The Neurophysiology of Language Processing Shapes the Evolution of Grammar: Evidence from Case Marking." *PlosONE* 10 (8): e0132819.

Bowern, Claire. 2008. "Bardi Arguments: Referentiality, Agreement, and Omission in Bardi Discourse." In *Discourse and Grammar in Australian Languages*, ed. by Ilana Mushin, and Brett Baker, 59–85. Amsterdam: John Benjamins.

Caudal, Patrick, and Robert Mailhammer. 2022. "Linear Lengthening in Iwaidja: An Event-Quantifying Intonation at the Phonology to Semantics/Pragmatics Interface." *Languages* 7 (3): 209.

Chadwick, Neil. 1997. "The Barkly and Jaminjung Languages: A Non-Contiguous Genetic Grouping." In *Boundary Rider: Studies in the Lexicology and Comparative Linguistics of Australian Languages*, ed. by Darrell Tryon, and Michael Walsh, 95–106. Canberra: Pacific Linguistics.

Chafe, Wallace. 1976. "Givenness, Contrastiveness, Definiteness, Subjects, Topics and Point of View." In *Subject and Topic*, ed. by Charles N. Li, 25–55. New York: Academic Press.

Dik, Simon C. 1997. *The Theory of Functional Grammar. Part 1: The Structure of the Clause*. Berlin: De Gruyter Mouton.

Dryer, Matthew S. 1986. "Primary Objects, Secondary Objects, and Antidative." *Language* 62 (4): 808–845.

Dryer, Matthew S. 2013a. "Chapter 81: Order of Subject, Object and Verb." In *The World Atlas of Language Structures Online*, ed. by Matthew S. Dryer, and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, Matthew S. 2013b. "Chapter 82: Order of Subject and Verb." In *The World Atlas of Language Structures Online*, ed. by Matthew S. Dryer, and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, Matthew S. 2013c. "Chapter 83: Order of Object and Verb." In *The World Atlas of Language Structures Online*, ed. by Matthew S. Dryer, and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, Matthew S. 2013d. "Determining Dominant Word Order." In *The World Atlas of Language Structures Online*, ed. by Matthew S. Dryer, and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Du Bois, John W. 1987. "The Discourse Basis of Ergativity." *Language* 63 (4): 805–855.

Du Bois, John W. 2003. "Argument Structure: Grammar in Use." In *Preferred Argument Structure: Grammar as Architecture for Function*, ed. by John W. Du Bois, Lorraine E. Kumpf, and William J. Ashby, 11–60. Amsterdam: John Benjamins.

Egurtzegi, Aitor, Sebastian Sauppe, Arrate Isasi-Isasmendi, Gillen Martinez de la Hidalga, Ina Bornkessel-Schlesewsky, Matthias Schlesewsky, Itziar Laka, Martin Meyer, Balthasar Bickel, and Caroline Andrew. 2025. "The Effect of Animacy on the Agent Preference: Self-Paced Reading Evidence from Basque." *Memory and Cognition* 53: 2056–2075.

Everett, Caleb. 2009. "A Reconsideration of the Motivations for Preferred Argument Structure." *Studies in Language* 33 (1): 1–24.

Gell-Mann, Murray, and Merritt Ruhlen. 2011. "The Origin and Evolution of Word Order." *Proceedings of the National Academy of Sciences* 108 (42): 17290–17295.

Gerdes, Kim, Sylvain Kahane, and Xinying Chen. 2021. "Typometrics: From Implicational to Quantitative Universals in Word Order Typology." *Glossa: A Journal of General Linguistics* 6 (1): 17.

Goldin-Meadow, Susan, Wing Chee So, Aslı Özyürek, and Carolyn Mylander. 2008. "The Natural Order of Events: How Speakers of Different Languages Represent Events Nonverbally." *Proceedings of the National Academy of Sciences* 105 (27): 9163–9168.

Green, Ian, and Rachel Nordlinger. 2004. "Revisiting Proto-Mirndi." In *Australian Languages: Classification and the Comparative Method*, ed. by Claire Bowern, and Harold Koch, 291–311. Amsterdam: John Benjamins.

Greenberg, Joseph H. 1966. "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements." In *Universals of Language: Report of a Conference Held at Dobbs Ferry, New York, April 13–15, 1961*, ed. by Joseph H. Greenberg, 73–113. Cambridge, MA: MIT Press.

Gregory, Michelle L., and Laura A. Michaelis. 2001. "Topicalization and Left-Dislocation: A Functional Opposition Revisited." *Journal of Pragmatics* 33: 1665–1706.

Grosz, Barbara, and Yael Ziv. 1997. "Centering, Global Focus, and Right Dislocation." In *Centering Theory in Discourse*, ed. by Marilyn A. Walker, Aravind Krishna Joshi, and Ellen F. Prince, 293–308. Oxford: Oxford University Press.

Haig, Geoffrey, and Stefan Schnell. 2014. "Annotations Using GRAID (Grammatical Relations and Animacy in Discourse): Manual Version 7.0." University of Bamberg. https://fis.uni-bamberg.de/handle/uniba/21446

Haig, Geoffrey, and Stefan Schnell. 2016. "The Discourse Basis of Ergativity Revisited." *Language* 92 (3): 591–618.

Haig, Geoffrey, Stefan Schnell, and Claudia Wegener. 2011. "Comparing Corpora from Endangered Language Projects: Explorations in Language Typology Based on Original Texts." In *Documenting Endangered Languages*, ed. by Geoffrey Haig, Nicole Nau, Stefan Schnell, and Claudia Wegener, 55–86. Berlin, Boston: De Gruyter Mouton.

Hall, Matthew L., Rachel I. Mayberry, and Victor S. Ferreira. 2013. "Cognitive Constraints on Constituent Order: Evidence from Elicited Pantomime." *Cognition* 129 (1): 1–17.

Harvey, Mark. 2008. *Proto Mirndi. A Discontinuous Language Family in Northern Australia*. Canberra: Pacific Linguistics.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2006a. "Language Documentation: What Is It and What Is It Good For?' In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2006b. "Prosody in Language Documentation." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 163–181. Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2006c. "The Challenges of Segmenting Spoken Language." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 253–274. Berlin: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2012. "Linguistic Data Types and the Interface between Language Documentation and Description." *Language Documentation and Conservation* 6: 187–207. http://hdl.handle.net/10125/4503

Himmelmann, Nikolaus P. 2016. "What about Typology is Useful for Language Documentation?' *Linguistic Typology* 20 (3): 473–478.

Himmelmann, Nikolaus P., and Beatrice Primus. 2015. "Prominence beyond Prosody." In *pS-prominenceS: Prominences in Linguistics*, ed. by Amedeo De Dominicis, 38–58. Viterbo: DISUCOM Press.

Hockett, Charles F. 1958. *A Course in Modern Linguistics*. New York: Macmillan.

Huber, Eva, Sebastian Sauppe, Arrate Isasi-Isasmendi, Ina Bornkessel-Schlesewsky, Paola Merlo, and Balthasar Bickel. 2024. "Surprisal from Language Models can Predict ERPs in Processing Predicate-Argument Structures only if Enriched by an Agent Preference Principle." *Neurobiology of Language* 5 (1): 167–200.

Hung, Yu-Chen, and Petra B. Schumacher. 2012. "Topicality Matters: Position-Specific Demands on Chinese Discourse Processing." *Neuroscience Letters* 511 (2): 59–64.

Jacobs, Joachim. 2001. "The Dimensions of Topic-Comment." *Linguistics* 39: 641–681.

Kaiser, Elsi, and John C. Trueswell. 2004. "The Role of Discourse Context in the Processing of a Flexible Word-Order Language." *Cognition* 94 (2): 113–147.

Koizumi, Masatoshi. 2023. *Constituent Order in Language and Thought: A Case Study in Field-Based Psycholinguistics*. Cambridge: Cambridge University Press.

Koizumi, Masatoshi, Yoshiho Yasugi, Katsuo Tamaoka, Sachiko Kiyama, Jungho Kim, Juan Esteban Ajsivinac Sian, and Lolmay Pedro Oscar García Mátzar. 2014. "On the (Non)Universality of the Preference for Subject-Object Word Order in Sentence Comprehension: A Sentence-Processing Study in Kaqchikel Maya." *Language* 90 (3): 722–736.

Krifka, Manfred. 2001. "Quantifying into Question Acts." *Natural Language Semantics* 9: 1–40.

Krifka, Manfred, and Renate Musan. 2012. "Information Structure: Overview and Linguistic Issues." In *The Expression of Information Structure*, ed. by Manfred Krifka, and Renate Musan, 1–44. Berlin: De Gruyter Mouton.

Kuroda, S.-Y. 1972. "The Categorical and the Thetic Judgment. Evidence from Japanese Syntax." *Foundations of Language* 9: 153–185. http://www.jstor.org/stable/25000656

Lambrecht, Knud. 1981. *Topic, Antitopic and Verb Agreement in Non-Standard French*. Amsterdam: John Benjamins.

Lambrecht, Knud. 1994. *Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.

Lambrecht, Knud. 2000. "When Subjects Behave like Objects: An Analysis of the Merging of S and O in Sentence-Focus Constructions across Languages." *Studies in Language* 24 (3): 611–682.

Lambrecht, Knud. 2001. "Dislocations." In *Language Typology and Language Universals: An International Handbook*, vol. 2, ed. by Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, 1050–1078. Berlin: De Gruyter Mouton.

Levshina, Natalia. 2019. "Token-Based Typology and Word Order Entropy: A Study Based on Universal Dependencies." *Linguistic Typology* 23 (3): 533–572.

Levshina, Natalia, Savithry Namboodiripad, Marc Allassonnière-Tang, et al. 2023. "Why We Need a Gradient Approach to Word Order." *Linguistics* 61 (4): 825–883.

Mayer, Mercer. 1969. *Frog, Where Are You?* New York: Dial Press.

Mithun, Marianne. 1992. "Is Basic Word Order Universal?' In *Pragmatics of Word Order Flexibility*, ed. by Doris Payne, 15–61. Amsterdam: John Benjamins.

Nichols, Johanna. 1986. "Head-Marking and Dependent-Marking Grammar." *Language* 62 (1): 56–119.

Nordlinger, Rachel, Gabriela Garrido Rodriguez, and Evan Kidd. 2022. "Sentence Planning and Production in Murrinhpatha, an Australian 'free Word Order' Language." *Language* 98 (2): 187–220.

O'Shannessy, Carmel. 2004. "The Monster Stories: Picture Stimuli for Elicited Production." Unpublished Manuscript. Nijmegen: Max Planck Institute for Psycholinguistics.

Ponsonnet, Maïa. 2023. "Possession." In *The Oxford Guide to Australian Languages*, ed. by Claire Bowern, 243–252. Oxford: Oxford University Press.

Portner, Paul, and Katsuhiko Yabushita. 1998. "The Semantics and Pragmatics of Topic Phrases." *Linguistics and Philosophy* 21 (2): 117–157.

Riesberg, Sonja, and Beatrice Primus. 2015. "Agent Prominence in Symmetrical Voice Languages." *STUF — Language Typology and Universals* 68 (4): 551–564.

Riesberg, Sonja, Kurt Malcher, and Nikolaus P. Himmelmann. 2019. "How Universal Is Agent-First? Evidence from Symmetrical Voice Languages." *Language* 95 (3): 523–561.

Sasse, Hans-Jürgen. 1987. "The Thetic/Categorical Distinction Revisited." *Linguistics* 25 (3): 511–580.

Sasse, Hans-Jürgen. 2006. "Theticity." In *Pragmatic Organization of Discourse in the Languages of Europe*, ed. by Giuliano Bernini, and Marcia L. Schwarz, 255–308. Berlin: De Gruyter Mouton.

Sauppe, Sebastian, Åshild Næss, Giovanni Roversi, Martin Meyer, Ina Bornkessel-Schlesewsky, and Balthasar Bickel. 2023. "An Agent-First Preference in a Patient-First Language During Sentence Comprehension." *Cognitive Science* 47 (9): e13340.

Schneider-Blum, Gertrud. 2023. "The Impact of Attentional Centering on Ergative Marking in Tima." *STUF — Language Typology and Universals* 76 (1): 87–112.

Schneider-Blum, Gertrud, and Birgit Hellwig. 2018. "Reference Tracking in Tima and Its Interplay with Split Ergative Marking." *Studies in Language* 42 (4): 970–993.

Schnell, Stefan, and Nils Norman Schiborr. 2022. "Crosslinguistic Corpus Studies in Linguistic Typology." *Annual Review of Linguistics* 8 (1): 171–191.

Schultze-Berndt, Eva. 2000. Simple and Complex Verbs in Jaminjung: A Study of Event Categorisation in an Australian Language. PhD thesis, University of Nijmegen.

Schultze-Berndt, Eva. 2003. "Preverbs as an Open Word Class in Northern Australian Languages: Synchronic and Diachronic Correlates." In *Yearbook of Morphology 2003*, ed. by Geert Booij, and Jaap van Marle, 145–177. Dordrecht: Kluwer.

Schultze-Berndt, Eva. 2010. "Grammatical Properties and Classification of Three-Participant Predicates in Jaminjung." In *Studies in Ditransitive Constructions*, ed. by Andrej Malchukov, Martin Haspelmath, and Bernard Comrie, 510–528. Berlin: De Gruyter Mouton.

Schultze-Berndt, Eva. 2012. "Pluractional Posing as Progressive: A Construction between Lexical and Grammatical Aspect." *Australian Journal of Linguistics* 32 (1): 7–39.

Schultze-Berndt, Eva. 2015. "Complex Verbs, Simple Alternations: Valency and Verb Classes in Jaminjung." In *Valency Classes in the World's Languages. Vol. 2: Case Studies from Austronesia, the Pacific, the Americas, and Theoretical Outlook*, ed. by Andrej Malchukov, and Bernard Comrie, 1117–1162. Berlin: De Gruyter Mouton.

Schultze-Berndt, Eva. 2017a. "Interaction of Ergativity and Information Structure in Jaminjung (Australia)." In *Oxford Handbook of Ergativity*, ed. by Jessica Coon, Diane Massam, and Lisa deMena Travis, 1089–1113. Oxford: Oxford University Press.

Schultze-Berndt, Eva. 2017b. "Two Classes of Verbs in Northern Australian Languages: Implications for the Typology of Polycategoriality." In *Lexical Polycategoriality: Cross-Linguistic, Cross-Theoretical and Language Acquisition Approaches*, ed. by Valentina Vapnarsky, and Edy Veneziano, 243–271. Amsterdam: John Benjamins.

Schultze-Berndt, Eva. 2022. "When Subjects Frame the Clause: Discontinuous Noun Phrases as an Iconic Strategy for Marking Thetic Constructions." *Linguistics* 60 (3): 865–898.

Schultze-Berndt, Eva. 2023. "The Reflexive Construction in Jaminjung/Ngaliwurru." In *Reflexive Constructions in the World's Languages*, ed. by Katarzyna Janic, Nicoletta Puddu, and Martin Haspelmath, 541–565. Berlin: Language Science Press.

Schultze-Berndt, Eva. 2025a. "Annotated Dataset for the Paper: Information Structure and the Agent-first preference: A Corpus-based Case Study of a Free Word Order Language." *Dataset*. University of Manchester.

Schultze-Berndt, Eva. 2025b [2020]. "GRAID + Information Structure: Annotation Decisions for Jaminjung-Ngaliwurru, July 2020." Manuscript. University of Manchester.

Schultze-Berndt, Eva, and Felicity Meakins. 2017. "Bringing Agents into Focus: Competing Motivations in Differential Agent Marking." Presentation, 12th International Meeting of the Association for Linguistic Typology, Canberra, 11–15 December 2017.

Schultze-Berndt, Eva, Stefan Schnell, and Ruth Singer. 2025 [2020]. "GRAID + Information Structure Annotations." *Manuscript*. University of Manchester.

Schultze-Berndt, Eva, and Candide Simard. 2012. "Constraints on Noun Phrase Discontinuity in an Australian Language: The Role of Prosody and Information Structure." *Linguistics* 50 (5): 1015–1058.

Schultze-Berndt, Eva, Candide Simard, Mark Harvey, Dorothea Hoffmann, and Jaminjung and Ngaliwurru communities. 2017. "Annotated Audiovisual Corpus of Jaminjung and Ngaliwurru, 1993–2016." DobeS Archive. https://dobes.mpi.nl/projects/jaminjung/

Sedarous, Yourdanis, and Savithry Namboodiripad. 2020. "Using Audio Stimuli in Acceptability Judgment Experiments." *Language and Linguistics Compass* 14 (8): e12377.

Simard, Candide. 2010. The Prosodic Contours of Jaminjung, a Language of Northern Australia. PhD thesis, University of Manchester. https://research.manchester.ac.uk/en/studentTheses/the-prosodic-contours-of-jaminjung-a-language-of-northern-austral/

Simard, Candide. 2013a. "Marking Boundaries: Intonation Units and Prosodic Sentences." In *Proceedings of Prosody-Discourse Interface Conference 2013 (IDP-2013)*, ed. by Piet Mertens, and Anne Catherine Simon, 131–136. https://www.arts.kuleuven.be/ling/cohistal/conference/idp2013/documents/proceedings_idp2013

Simard, Candide. 2013b. "Prosodic Encoding of Declarative, Interrogative and Imperative Sentences in Jaminjung, a Language of Australia." *Proceedings of Interspeech 2013*: 793–797.

Simard, Candide. 2014. "Another Look at Right Detached NPs." In *Proceedings of Language Documentation and Linguistic Theory 4*, ed. by Aicha Belkadi, Kakia Chatsiou, and Kirsty Rowan. London: SOAS. https://www.elpublishing.org/docs/11/4/ldlt4_17.pdf

Simard, Candide, and Eva Schultze-Berndt. 2011. "Documentary Linguistics and Prosodic Evidence for the Syntax of Spoken Language." In *Documenting Endangered Languages: Achievements and Perspectives*, ed. by Geoffrey Haig, Nicole Nau, Stefan Schnell, and Claudia Wegener, 151–176. Berlin: De Gruyter Mouton.

Sinnemäki, Kaius. 2010. "Word Order in Zero-Marking Languages." *Studies in Language* 34 (4): 869–912.

Tomlin, Russell S. 1986. *Basic Word Order*. London: Routledge.

Wälchli, Bernhard. 2009. "Data Reduction Typology and the Bimodal Distribution Bias." *Linguistic Typology* 13 (1): 77–94.

Whaley, Lindsay. 1997. *An Introduction to Typology: The Unity and Diversity of Language*. Thousand Oaks: Sage Publications.

Ziv, Yael. 1994. "Left and Right Dislocations: Discourse Functions and Anaphora." *Journal of Pragmatics* 22 (6): 629–645.

CHAPTER 25

Bridging demonstratives in recognitional use

Petra B. Schumacher & Klaus von Heusinger
University of Cologne

Reference to an implicit argument in a bridging context is made by definite determiner phrases (DPs) (*a book... the author*), and it is assumed that demonstrative DPs are not felicitous in such contexts (*a book... #that author*). However, the literature reports bridging contexts with demonstrative DPs. We argue that the recognitional use of demonstratives, first described in Himmelmann (1996, 1997), can license demonstrative DPs in such contexts. Bridging contexts rely on general knowledge, making definite DPs the default in the "larger situational use" (Hawkins 1978). We argue that demonstrative DPs are possible in such contexts if they signal special shared knowledge between speaker and hearer, i.e. if they are recognitional. Two rating tasks with evaluative adjectives provide evidence for the recognitional function of bridging demonstratives.

Keywords: demonstrative, recognitional use, bridging, definite article, evaluative adjective, metalinguistic judgement, acceptability task

1. Introduction

Demonstrative expressions exist in all languages (Himmelmann 1997; Diessel 1999) and have different uses. In the following, we focus on demonstrative determiner phrases (DPs, i.e. the combination of a demonstrative determiner and a noun, as in *this author*). In addition to the three uses discussed in prior literature, namely (i) the deictic, (ii) the discourse deictic, and (iii) the anaphoric or (tracking) use, Himmelmann (1996, 1997) introduced a fourth use, the recognitional use, which is characterized by special and shared knowledge between the speech participants. Himmelmann also uses this special context to capture the grammaticalization from a demonstrative function to the definite article function. A key difference between the use of a definite DP and a demonstrative DP is that the definite article is used when the intended referent is uniquely identifiable based

https://doi.org/10.1075/slcs.240.25sch
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

on encyclopaedic knowledge or Hawkin's (1978) "larger situational use" (*the sun*), where the demonstrative is not appropriate in such contexts (*#this sun*). This contrast also holds for so-called bridging contexts, where the context makes available an anchor expression for a discourse-new inferred referent that is unique and can be captured with a definite DP (*Phoebe read a book. The author…*), but not with a demonstrative (*Phoebe read a book. #This author…*). Thus, bridging contexts are often used as test cases in languages that grammaticalize a definite article from the demonstrative, in order to find out whether the "demonstrative" expression has already assumed the function of a definite article (see also Himmelmann 1996).

At the same time, the literature occasionally provides counterexamples to the generalization that demonstrative DPs are not allowed in bridging contexts (Gundel et al. 1993; Apothéloz & Reichler-Béguelin 1999; Schwarz 2000; Wolter 2006; Doran 2012; Doran & Ward 2019). Most of the examples are accounted for by certain additional parameters, such as contrast or avoidance of genericity. Doran (2012: 124) argues that demonstrative DPs can be used if they express a contrast. Accordingly, the demonstrative in (2) is felicitous, but not in (1) (from Doran 2012: 124–125).

(1) *Sorry I'm late. I grabbed a taxi but {the/#that} driver took his time.*

(2) *My taxi was going too slow so I jumped out and hailed another. That {driver/ second driver} went even slower.*

However, there are examples where no additional parameter such as contrast seems to be present, as in (3) from Gundel et al. (1993: 282).

(3) *We went to hear the Minnesota Orchestra last night.* ***That conductor*** *was very good.*

The use of the demonstrative DP instead of the expected definite DP needs some explanation. We argue in this paper that a demonstrative DP can be licensed in bridging contexts when it is used in its recognitional function. We think that such a recognitional use is triggered by invoking the thoughts and feelings of a speaker with respect to a particular entity. This can be accomplished by using evaluative adjectives, such as *arrogant* in (4). The speaker refers to the journalist responsible for the interview (through the conceptual bridging relation between interview and journalist) and further suggests that the journalist in question is already known by the addressee (via the use of the demonstrative determiner in combination with the evaluative adjective).

(4) *Stefan heard an interview with NASA yesterday.* ***That arrogant journalist*** *had to do a lot of research for it.*

The remainder of this paper proceeds as follows: In Section 2, we define the central concepts — (a) demonstrative vs. definite articles, (b) recognitional use of demonstratives, (c) indirect or bridging anaphora — and we describe the relationship between these concepts. Section 3 then describes two rating tasks that probe speakers' metalinguistic awareness as a case of language documentation in the sense of Himmelmann (1998), and Section 4 summarizes the results and gives an outlook.

2. Demonstratives and inferable referents

2.1 Demonstrative and definite articles

Demonstratives are typically characterized by expressing a contrast with some other form in terms of distance to the speaker ("proximal" vs. "distal") or by the function to raise the attention of the hearer towards a less accessible entity. In fact, in many languages, both of these features are associated with demonstratives, which are typically used as pronouns (*this*) or as determiners with noun phrases (*this book*) (Diessel 1999; Dryer 2013). Definite articles, on the other hand, do not express physical distance from the speaker, they can only be used with a noun phrase, but not as a pronoun. Definite articles are used to signal to the hearer that the intended referent can be unambiguously identified, typically if it is unique (*the sun*) or familiar as in the anaphoric use (*I bought a book and a journal. The book was more expensive than the journal*).[1]

Thus, we assume that demonstrative DPs are clearly distinguishable from definite DPs. Demonstrative DPs can signal locational information, they raise the level of attention, they refer directly to their referents and they are not licensed in unique contexts (*#this sun*). Definite DPs do not convey locational information, they identify their referent by their descriptive material and they are licensed in unique context. There are two contexts that allow definite DPs, but not demonstrative DPs: abstract situational or unique uses (*the sun*) and bridging uses, as in (1) (Himmelmann 1997: 191; Becker 2021: 110). Both types can be used in (coreferential) anaphoric contexts, as in (5a)–(b). They show, however, an interesting contrast in contexts with two descriptive antecedents, as in (6a)–(b): While definite DPs felicitously refer back to the two antecedents, demonstrative DPs are not possible or at most marginally acceptable (see Hinterwimmer 2019).

1. A reviewer pointed out to us that Becker (2021: 99) provides further types of articles including anaphoric, weak and recognitional articles. We will use "definite article" for expressions that include at least anaphoric use, unique use and bridging use.

(5) a. *Ann owns a cat. The cat is a good companion.*
b. *Ann owns a cat. That cat is a good companion.*

(6) a. *Ann owns a cat and a dog. Usually, the cat chases the dog.*
b. *Ann owns a cat and a dog. #Usually, that cat chases that dog.*

2.2 Uses of demonstratives

Himmelmann (1996: 240), following the literature, describes three different uses of demonstratives: (i) the situation or deictic use, as in (7), where the intended referent is in the perceivable situation and optionally combined with a pointing gesture; (ii) the discourse deictic use, as in (8), where the referent points to parts of the previous discourse; and (iii) the tracking or anaphoric use, where the referent is discourse-given and explicitly accessible from prior discourse (9). Besides these core functions of demonstratives, which can be found in most languages (Diessel 1999; Doran & Ward 2019), Himmelmann (1996, 1997) adds a new use, namely the recognitional use, as in (10), which is based on special (and "personalized") knowledge of speaker and hearer (and which will be described in the next section). Further functions of demonstratives are the indefinite use, as in (11), where a demonstrative DP introduces a discourse- and hearer-new referent about which the speaker will provide more information (Prince 1981a). The use of demonstratives can also have an emphatic function, as in (12) (Lakoff 1974).

(7) *You can take ☞* ***this book****.*

(8) *He said "I love you" and with* ***these words*** *he left.*

(9) *Once upon a time there was a princess.* ***This princess*** *had a crown.*

(10) *Do you know what happened to* ***that phone*** *that used to be in your room?*

(11) *Yesterday I walked into a bar and there was* ***this stranger*** *who stared at me all the time.*

(12) *And then I met* ***this neighbour*** *of yours.*

Table 1 provides an overview of these six functions and describes the discourse status and further features (von Heusinger 2011).[2]

2. There are further uses of demonstratives, such as generic or self-reference uses etc. (Levinson 2004; Doran & Ward 2019; Maes et al. 2022).

Table 1. Uses of demonstratives

Demonstrative uses	Discourse-given	Speaker-known	Hearer-known	Further characterization	Ex.
deictic	✓	✓	✓	perceivable in situation	(7)
discourse deictic	✓	✓	✓	reference to part of discourse	(8)
anaphoric	✓	✓	✓	reference to discourse entity	(9)
recognitional	✗	✓	✓	shared (private) knowledge	(10)
indefinite	✗	✓	✗	speaker's knowledge	(11)
emphatic	✗	✓	✗	emotional / social relation	(12)

2.3 Recognitional use

Himmelmann (1996, 1997) introduced the "recognitional use" for first-mentioned, discourse-new demonstratives that cannot be resolved with respect to the immediate situation or via an anaphoric link, but rather to a specific knowledge that is shared by speaker and hearer:

> Recognitional use of demonstratives, on the other hand, draws on specific, "personalized" knowledge that is assumed to be shared by the communicating parties due to a common interactional history or to supposedly shared experiences.
> (Himmelmann 1996: 233)[3]

Himmelmann (1997: 57) illustrates this use with *diese Tiefkühltruhen* ('those freezers') in (13), which is licensed because it refers to an assumed privately shared knowledge of writer and reader. In (14) from the Pear Stories (Himmelmann 1996: 230), the use of the demonstrative DP *those dusty kinds of hills* signals that the speaker assumes that the hearer has the same specific knowledge about the

3. In addition to the main characteristics (i), Himmelmann (1997: 81) provides two further features for the recognitional (his "anamnestisch") use of demonstratives (modifying the description of Auer 1984):

i. first-mentioned referent such that both speaker and hearer have common specialized knowledge that allows to identify the referent;
ii. speaker can also express uncertainty whether there is such a shared special knowledge, speaker signals that they will provide more information if requested;
iii. speaker can also signal that they are not certain about the correct referential expression.

intended referent. A possible diagnostic for recognitional use is that it can be accompanied by tag questions such as "you know?" or "remember?" (Himmelmann 1996: 230).

(13) *Die soll so kochen, wie ihre Mutter oder Großmutter gekocht hat. Die hatte das gar nicht alles zur Verfügung — **diese Tiefkühltruhen** und den Quatsch.*
[Kölner Stadtanzeiger 238/12.10.93:4, Interview mit Siebeck]
'She should cook the way her mother or grandmother cooked. She didn't have all that at her disposal — those freezers and all that nonsense.'

(14) *It was filmed in California, those dusty kinds of hills that they have out here by Stockton and...*

Himmelmann (1997: 62–72) provides additional data from the Australian language Nunggubuyu that has four types of demonstratives (proximal, medial, distal, anaphorical) and he argues that speakers use the "anaphorical" demonstrative in recognitional contexts. This use is typically found in mythological stories where speakers can assume that many characters are known and need not be introduced explicitly. But the recognitional use can also be found in everyday life, as illustrated by Himmelmann by the following example from Heath (1980: 161–162).

> Often I would head into the Aboriginal part of the village ... obviously in search of my regular informant. If he was not at home, someone else there would say to me before I could open my mouth:
> *ni-ya-nggi* *bu-gu-ni* *nu-ba-gi-yung*
> 3.SG.M=go-PAST2 ANAPH-LOC.ADV-ALL M.SG:ANAPH-SG-ABS
> 'That one went there.'

Himmelmann (1997: 71) notes that such a recognitional use can be found cross-linguistically, even though it is often not explicitly listed in grammars, as they rather focus on locational information. For instance, Khachaturyan (2019: 199) reports different types of recognitional uses from Mano (Niger-Congo), spoken in Guinea. The demonstrative *yā* signals that speaker and hearer have shared knowledge about the task to go to the field on the next day.

(15) *tòò* *yā,* *kóò* *lō* *bú* *yā* *mè-è.*
tomorrow DEM 1PL.IPFV go:IPFV rice:FOC DEM beat-GER
'THAT tomorrow, we will beat THAT rice.'

Khachaturyan also presents examples where the recognitional demonstrative is used in religious texts, through which the speaker suggests that the referent is identifiable. She argues that a recognitional demonstrative can be used when both speaker and hearer have shared knowledge about the intended referent or it can

be used by the speaker to indicate to the hearer to accommodate the referent based on shared knowledge.

Besides the term "recognitional use", we find alternative terms: Gundel et al. (1993: 302) use the term "reminder *that*" for the observation that the use of a demonstrative *that* for a familiar referent implies that the referent is not activated, since otherwise the unmarked definite article would have been used:

(16) *Exxon Oil claims it will take several million dollars to clean up* ***that oil spill off the coast of Alaska.*** [beginning of radio newscast]

Himmelmann's term "recognitional use" is now the common term for this kind of use (Levinson 2004; Doran & Ward 2019; Maes et al. 2022). According to Himmelmann (1997: 61), the demonstrative DP is justified because it creates a situation in which speaker and hearer have direct access to the referent. This is clearly different from the use of the definite article for the larger situational use according to Hawkins, as in (17).

(17) *The sun was about to approach its zenith.*

For Himmelmann, the recognitional use is the "missing link" in the grammaticalization process from demonstratives to definite articles. He assumes that the demonstrative grammaticalizes from its anaphoric function to the recognitional function. The next grammaticalization step is that the form can be used not only when speaker and hearer have special shared knowledge, but also when they share general knowledge. If we find such a use of the form, we can assume that the demonstrative has the function of the definite article (see also de Mulder & Carlier 2011; Becker 2021: 108–109).

In sum, a central characteristic of recognitional use is that it is interactional and relies on shared knowledge that is specific to the interlocutors. The corresponding referent must be retrieved from representations that are based on shared experience and the speaker assumes that the addressee can access the relevant representation from collectively acquired knowledge. By using a demonstrative expression, the speaker indicates that discourse context and the situation of utterance do not suffice for reference identification and that additional information has to be activated from shared knowledge (Auer 1984). The pointer to shared knowledge is sometimes reflected by translations such as "the one you should remember" or "you know the one" (Goddard 1983; Wilkins 1989). This function is further qualified by the observation that some languages make available a dedicated form for this use (Himmelmann 1996).

Further below, we operationalize shared knowledge through the use of evaluative expressions, which we do not associate with the emphatic function of demonstratives. A reviewer pointed out that Levinson (2004: 108) draws a close relation

between the recognitional use and the emphatic (his "empathetic") use, originally introduced by Lakoff (1974), who refers to one type of "emotional deixis" as the "solidarity type" creating a group feeling. Chen (1990: 148–151) lists three types of emotional uses of demonstratives: (i) distancing-*that*, (ii) sympathy-*that*, and (iii) camaraderie-*that*. The solidarity type of Lakoff and the camaraderie type of Chen share some features with the recognitional use described above, in particular if we assume with Khachaturyan (2019) that a recognitional use can also be licensed if the hearer is able to accommodate the inferred referent and thus create that shared knowledge. Doran (2012: 94) argues, however, that the emphatic or emotional aspect of demonstratives is not an independent use, but an additional pragmatic feature of a demonstrative if the referent can be identified by speaker and hearer. So, the emphatic aspect can be added to deictic as well as anaphoric demonstratives. Emphatic demonstratives are typically accented, while recognitional demonstrative are deaccented. We therefore assume that the recognitional use is different from the emphatic one.

2.4 Bridging inferences

Definite DPs typically have a linguistic antecedent, as *the cat* and *the dog* in (5) and (6) above, or they can refer to unique entities according to our world knowledge, as in *the sun* in (17). However, we also find many first-mentioned definite DPs that do not fulfil these requirements (Fraurud 1990; Poesio & Veira 1998). Their referents must be inferred from the content of the preceding discourse and accordingly, such definite expressions are called "indirect reference by association" (Clark 1975), "associative anaphora" (Hawkins 1978), "inferrables" (Prince 1981b) or "indirect anaphora" (Schwarz 2000). Establishing reference of such discourse-new entities exerts costs during processing (Burkhardt 2006). It is generally assumed that only definite DPs can be used to signal such indirect reference, but not demonstrative DPs or pronouns as in (18; slightly modified from Hawkins 1978: 87).[4]

(18) *Ann was discussing an interesting book in her class.*
  a. *She is friendly with the author.*
  b. *She is friendly with #this author.*
  c. *She is friendly with #him.*

4. It is interesting that bridging relations are possible in the deictic use, then called 'deferred reference' (Nunberg 1993). One can easily point to a picture and say "This [☞ to a picture] painter is my favorite one".

Hawkins (1978: 123) illustrates on the basis of (19) and (20) the conditions according to which a bridging definite is licensed. The definite DP *the dog* in (19) is not felicitous, as the anchors available are insufficient, i.e. we do not have general or encyclopaedic knowledge that each car or house has a dog. But we do have knowledge that a car exhausts fumes (if it is not an electric car) and therefore the definite DP *the exhaust fumes* in (20) is felicitous.

(19) *The man drove past our house in a car. #The dog was barking furiously.*

(20) *The man drove past our house in a car. The exhaust fumes were terrible.*

As indicated above, Himmelmann (1996: 210) states that demonstrative DPs should not emerge in larger situational use for unique referents (*this sun* vs. (17)) or for inferable referents (18b) (see also Gundel et al. 2000). However, there are some exceptions for demonstrative DPs. In (21), a corpus item from a fertility website (Doran & Ward 2019: 242), the demonstrative DP *those two double lines* is not given in the previous discourse, but it is inferable because the context introduces the framework of a pregnancy test.

(21) *I was told that when I conceive again I will need to take an extra 4 mg. folio acid daily along with 81 mg. of aspirin. Does anyone feel like they should take this prior to conceiving? Or wait till you see those two double lines?*

Wolter (2006: 6) provides (22), where the demonstrative DP is licensed because it forms a contrast relation with the initially introduced horn. In turn, without such a contrast, the demonstrative DP is not felicitous, as illustrated in (23).

(22) *A car drove by. The horn was honking. Then another car drove by. That horn was honking even louder.*

(23) *A car drove by. #That horn was honking.*

Wolter (2006: 9) attributes such cases to the shifting function of demonstratives. She assumes that "[d]emonstrative descriptions refer uniquely relative to a shifted domain" In this way, she can account for the contrast in (22)–(23), but not for the bridging inference in (21). We propose that the use of demonstratives in bridging contexts such as (21) can be better explained by assuming that they have a recognitional function.

2.5 Demonstratives in bridging context as recognitional use

We have seen in the last section that contrary to common assumptions in the literature, demonstratives are acceptable in bridging contexts under certain conditions. Bridging contexts typically license definite DPs, but not pronouns or demonstrative DPs, since definite DPs allow for both, descriptive inferences and the uniqueness condition. While the definite article is typically less marked and more frequently used in such contexts, demonstratives in bridging contexts serve a distinct discourse function best described as "recognitional". In this use, a demonstrative signals reference to a specific common ground shared between speaker and listener — knowledge that is more specific than the general encyclopaedic knowledge that would license the use of a definite DP. Bridging contexts, which normally license definite DPs due to the unique identifiability of referents based on shared world knowledge, provide for the comparison between the default use of definite DPs and the marked, perspective-driven use of demonstratives. In Example (3), repeated as (24), the recognitional use of the demonstrative DP *that conductor* creates or accommodates a special shared knowledge with potentially "personalized" knowledge of the particular individual, e.g. both speaker and hearer have heard the conductor before or had a discussion about him, etc.

(24) *We went to hear the Minnesota Orchestra last night.* ***That conductor*** *was very good.*

In the following, we focus on a specific bridging context, in which an event or product serves as anchor for a salient participant (e.g. lecture — professor, painting — artist). The inferred referent is typically presented with a definite determiner, but we argue that the use of a demonstrative leads to the activation of knowledge shared by the interlocutors. To reinforce the sharing of private knowledge, we also include evaluative adjectives, as in (25), which are understood to highlight the presence of a perspectival centre (Harris & Potts 2009) and reflect a comment of the speaker that alludes to shared knowledge and guides the addressee's attention to this personalized knowledge. Hence, evaluative adjectives represent a window to shared knowledge.

(25) *Stefan heard an interview with NASA yesterday.* ***That arrogant journalist*** *had to do a lot of research for it.*

To empirically test the function of recognitional demonstrative DPs in bridging contexts, we examine the interaction between determiner type (definite vs demonstrative) and evaluative adjectives (presence vs absence), which serve the purpose to point to knowledge shared by speaker and hearer.

3. Demonstratives in bridging context: Metalinguistic judgments

According to Himmelmann (1998), language documentation comprises the observation of linguistic behaviour in interaction *and* the recording of native speakers' metalinguistic judgments and interpretations. Metalinguistic knowledge emerges in comments on language as well as in playful use of language and it can be elicited in controlled experiments where members of a speech community are for instance asked to rate linguistic choices. In the following, we rely on this kind of metalinguistic knowledge to document the use conditions of demonstratives.

The aim of the current study was to examine under what conditions demonstrative DPs (DemDPs) are licensed in bridging contexts. In experiment 1, we tested the hypothesis that the presence of a perspective holder — implemented by the use of evaluative adjectives — boosts the acceptability of DemDPs. In experiment 2, we assessed whether the presence of an adjective has an effect on the referential choice.

3.1 Experiment 1

We hypothesize that one condition under which DemDPs are licensed in bridging contexts is when they are used for attentional orienting by a perspective holder in recognitional use. The presence of a perspectival anchor is implemented through an evaluative expression which contains a DP modified by an evaluate adjective (*that arrogant journalist*) and this condition is compared to an utterance with a neutral adjective (*that Italian journalist*).[5] DemDPs (26) are contrasted with definite DPs (in the following DPs) as in (27).

(26) a. DemDP-Evaluative: *Stefan hörte gestern ein Interview mit der NASA.* ***Dieser überhebliche Reporter*** *musste dafür sehr lange recherchieren.*
b. DemDP-Neutral: *Stefan hörte gestern ein Interview mit der NASA.* ***Dieser italienische Reporter*** *musste dafür sehr lange recherchieren.*

5. We use evaluative adjectives as a means to trigger the activation of shared knowledge in our experimental items that cannot rely on actual common ground between the speaker and the participants in the rating task. A reviewer suggested that the use of evaluative adjectives implies an egocentric perspective, while the recognitional use rather needs a sociocentric space (Skilton & Peeters 2021). The use of an evaluative adjective might combine these two aspects: On the one hand, the evaluative adjective must be anchored to the holder of the evaluation, and on the other hand, the evaluative adjective serves the function of pointing to the common ground shared by the interlocutors in our experimental vignettes, which do not have an elaborate context.

(27) a. DP-Evaluative: *Stefan hörte gestern ein Interview mit der NASA.* ***Der überhebliche Reporter*** *musste dafür sehr lange recherchieren.*
b. DP-Neutral: *Stefan hörte gestern ein Interview mit der NASA.* ***Der italienische Reporter*** *musste dafür sehr lange recherchieren.*
'Stefan heard an interview with NASA yesterday. That/The arrogant/Italian journalist had to do a lot of research for it.'

Using a 7-point acceptability task, we predicted (i) that DPs should receive higher ratings than DemDPs and (ii) that DemDPs would be rated better in the context of an evaluative (vs. neutral) adjective due to the licensing power of a perspectival anchor, while the rating of DPs should not be affected by the type of adjective.

We constructed 48 quadruplets with the factors DP type and adjective type, which were distributed across 4 lists in a Latin-square design and interspersed with 24 filler items (12 acceptable items without bridging relations and 12 inacceptable items with mismatching pronouns or incoherent continuations). The order of presentation of the items in a list was randomized for each participant. Native speakers of German ($n = 48$) completed the task remotely using the Qualtrics survey platform (Qualtrics, Provo, UT). They were asked to read each text carefully and rate how good each scenario sounded on a scale from 1 *sehr seltsam* ('very strange') to 7 *perfekt* ('perfect'). They indicated their response by clicking on stars below each item.

Ratings (1–7) for each item (including fillers) were converted to z-scores to account for variation in the participants' use of the scale. To establish that the task was performed appropriately, mean z-scores for the filler items were calculated and participants who rated unacceptable fillers higher than 3.5 or acceptable fillers lower than 3.5 were identified. This resulted in the exclusion of 12 participants on the basis of their ratings for the unacceptable fillers leaving 36 participants for the analysis. Participants' z-scores for the experimental items were analysed using linear mixed-effects models. The model contained the fixed factors DP type (DemDP, DP) and adjective type (evaluative, neutral) and random intercepts for participant and item; the inclusion of random slopes for the fixed factors were determined by assessment of singularity and by using the rePCA function in the RePsychLing package (Baayen et al. 2015).

Definite DPs were rated better than demonstrative DPs, but on a 7-point scale the latter were still considered acceptable. Evaluative adjectives had a positive effect on the ratings of DemDPs and lowered the ratings for definite DPs. Mean ratings and standard deviations (in parentheses) were 4.24 (1.55) for evaluative DemDPs, 4.11 (1.57) for neutral DemDPs, 5.17 (1.54) for evaluative DPs, 5.35 (1.38) for neutral DPs. Figure 1 shows the z-scores per condition.

The analysis showed an effect of DP type ($\beta = 0.46$, $SE = 0.072$, $t(62.84) = 6.46$, $p < .001$) and an interaction of DP type and adjective type ($\beta = 0.13$, $SE = 0.063$,

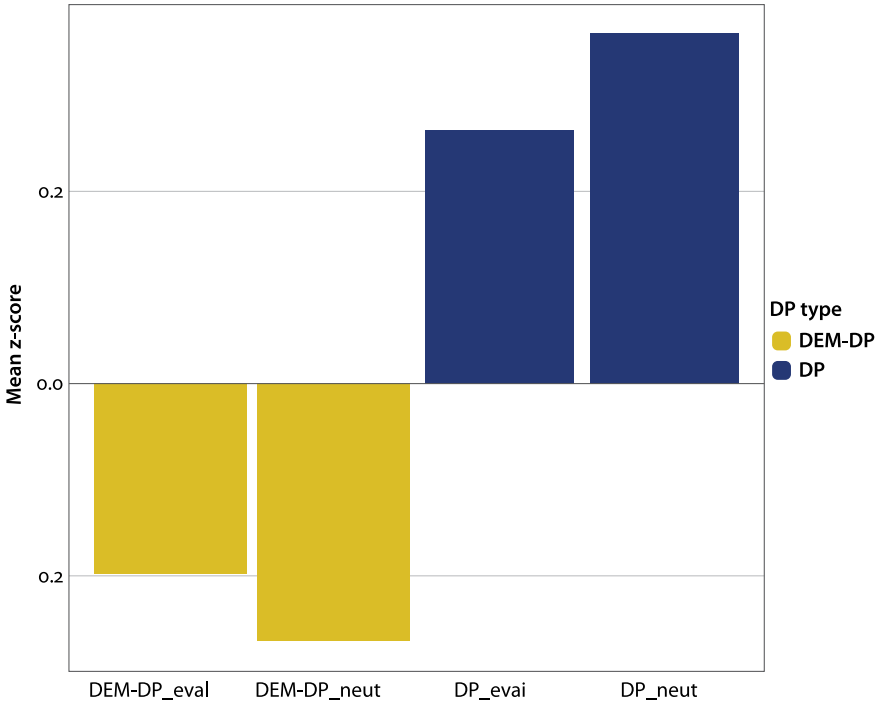


Figure 1. Mean z-scores per DP type and adjective type for experiment 1

$t(1546.59) = 2.15$, $p < .032$). Comparison of the DemDPs registered no reliable effect of adjective type ($p > .25$), but adjective type had an effect on the rating of DPs ($\beta = 0.09$, $SE = 0.044$, $t(785.99) = 1.98$, $p < .048$).

These data first confirm that definite DPs are licit referential forms in bridging contexts and that they receive higher ratings than demonstrative DPs. Interestingly, demonstrative DPs fare relatively well and receive a mean rating of 4.18. This counters the claim that demonstrative DPs are not possible in bridging contexts. The data further suggest that the two types of DPs are impacted differently by adjective type: while definite DPs benefit from the use of a neutral adjective, demonstrative DPs show higher ratings in combination with an evaluative adjective. This latter observation, although statistically only a trend, could be explained by the presence of a perspective holder who conveys an evaluation, which we claim contributes to the recognitional use and activates shared knowledge. The speaker's intention places an extra attentional focus on the referent and the attentional reorienting further benefits from the choice of a demonstrative determiner since demonstratives serve as pointers (here, to private knowledge). This may in turn also explain why the definite DPs show a reversed pattern with better ratings for the neutral than for the evaluative context: evaluative meanings may prefer to pair up with a demonstrative.

Given that the DemDPs are rated relatively high, an alternative account would be that the presence of an adjective renders these referential forms more acceptable in bridging contexts. To link our research to previous claims on the unavailability of DemDPs in bridging contexts, we carried out experiment 2 and contrasted DPs with and without modifying adjectives.

3.2 Experiment 2

We examined whether the presence or absence of a neutral adjective affects the ratings of definite and demonstrative DPs. Previous characterizations of the distribution of DemDPs suggest that bare DemDPs are poor referential forms in bridging contexts that do not communicate additional meaning aspects like contrast or attentional orienting (Hawkins 1978). Accordingly, they should not benefit from the presence or absence of a neutral adjective. Alternatively, the use of an adjective may make the referent more informative, drawing the attention on it in a way that DemDPs with adjectives are preferred over bare DemDPs. To test these predictions, we contrasted DemDPs (28) and DPs (29) with no adjective and with a neutral adjective:

(28) a. DemDP-Zero: *Stefan hörte gestern ein Interview mit der NASA.* ***Dieser Reporter*** *musste dafür sehr lange recherchieren.*
b. DemDP-Neutral: *Stefan hörte gestern ein Interview mit der NASA.* ***Dieser italienische Reporter*** *musste dafür sehr lange recherchieren.*

(29) a. DP-Zero: *Stefan hörte gestern ein Interview mit der NASA.* ***Der Reporter*** *musste dafür sehr lange recherchieren.*
b. DP-Neutral: *Stefan hörte gestern ein Interview mit der NASA.* ***Der italienische Reporter*** *musste dafür sehr lange recherchieren.*
'Stefan heard an interview with NASA yesterday. That/The Ø/Italian journalist had to do a lot of research for it.'

The experimental design and analyses were identical to experiment 1 with the exception that the stimuli with evaluative adjectives were replaced by DPs without an adjective. Out of the 48 native speakers of German who participated in the rating, ten were discarded from the analysis due to poor ratings for the unacceptable fillers. The remaining 38 participants' z-scores for the experimental items were analysed using linear mixed-effects models with the fixed factors DP type (DemDP, DP) and adjective (zero, neutral) as well as random intercepts for participant and item.

Figure 2 illustrates that definite DPs were again rated better than demonstrative DPs, but as before DemDPs still showed relatively high ratings. DPs without an adjective were rated higher than DPs with a neutral adjective, while DemDPs

showed no difference as a function of presence or absence of adjective. The conditions registered mean ratings and standard deviations (in parentheses) of 4.09 (1.47) for neutral DemDPs, 4.10 (1.57) for (zero) DemDPs, 5.45 (1.42) for neutral DPs and 5.77 (1.37) for (zero) DPs.

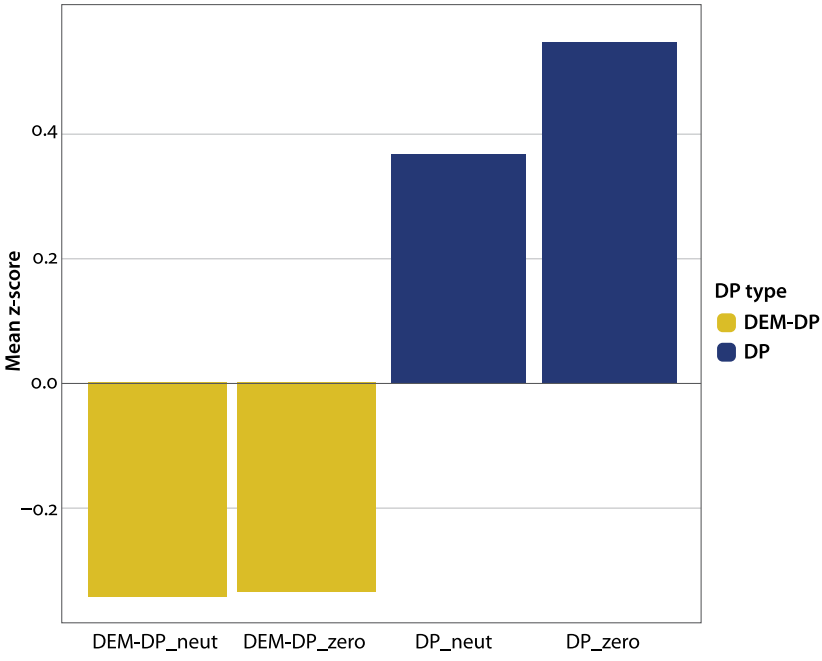


Figure 2. Mean z-scores per DP type and presence or absence of adjective for experiment 2

The linear mixed-effects analysis revealed an effect of DP type ($\beta = 0.71$, $SE = 0.089$, $t(78.20) = 7.96$, $p < .001$) and an interaction of DP type and adjective ($\beta = 0.16$, $SE = 0.054$, $t(1668.0) = 2.88$, $p < .004$). Comparison of the DemDPs registered no reliable effect of adjective ($p > .99$), but absence of an adjective made a difference for DPs ($\beta = 0.16$, $SE = 0.037$, $t(832.63) = 4.22$, $p < .001$).

The results of experiment 2 indicate that presence or absence of adjectives does not influence the acceptability ratings of demonstrative DPs. This suggests that the descriptively better ratings for DemDPs with evaluative adjectives observed in experiment 1 are closely tied to the evaluative meaning contribution. Definite DPs in bridging contexts seem to prefer as few extra information as possible, with DPs without modifications receiving the highest rating followed by neutral adjectives and evaluative DPs taking the poorest ratings.

3.3 Discussion

We tested the hypothesis that demonstrative DPs in inferential bridging contexts are one instance of the recognitional use and point to personalized knowledge shared by the interlocutors following Himmelmann's (1996) definition of this particular demonstrative function. Both experiments confirmed the view presented in the literature, that definite DPs are the preferred referential choice in bridging contexts. Yet interestingly, demonstrative DPs received relatively positive ratings, indicating that their use is not strictly unacceptable (contra the standard claim). Concerning the recognitional use, it was implemented by means of evaluative adjectives to accomplish perspectival anchoring and through this allude to mutual common ground in experiment 1. While evaluative adjectives — in contrast to neutral adjectives — had a positive impact on demonstrative DPs (i.e. higher ratings for evaluative DemDPs than neutral DemDPs), the effect was reversed for definite DPs (i.e. less acceptable scores for evaluative over neutral DPs). To follow up on the contribution of modifying adjectives on the ratings, we contrasted DPs with neutral adjectives with DPs without adjectival modifications in experiment 2 and found better ratings of unmodified definite DPs relative to definite DPs that contained a neutral adjective and no effect of the presence of (neutral) adjectives on demonstrative DPs. We conclude from this that demonstrative DPs benefit from the presence of an evaluative adjective, which we argue makes available the thoughts and evaluations of a perspectival anchor and triggers the use of personalized shared knowledge. Given that demonstrative DPs are rated best when accompanied by evaluative content, the demonstratives seem to serve the purpose of activating joint memory representations during recognitional use as suggested by Himmelmann (1996, 1997).

One reason for the marginal benefit of perspectival anchoring on the demonstrative DPs may be that recognitional use relies heavily on the interaction between speaker and hearer and marks the speaker's effort to make the addressee activate shared knowledge. In the present case, participants only read the stimuli, and using demonstrative DPs in more direct communicative situations, like dialogue, may very well improve the overall advantage of evaluative expressions on demonstrative use. In light of the documentarist turn, metalinguistic judgment tasks come with methodological limitations, but they represent crucial tasks for addressing subtle distinctions.

The definite DPs presented an interesting pattern across the two experiments that was not predicted. There were reliable differences between the three conditions tested, with unmodified definite DPs scoring highest, followed by neutral expressions and the lowest average rating for evaluative expressions. It seems to suggest that informativity has a negative effect on the rating of the stimuli with

definite expressions. The bridging relation may be more easily retrievable without interference from meaning components that are irrelevant to the establishment of a link between the anaphor and its anchor in this case; for instance, in the context of an interview, a journalist fills a necessary role that is easily accessible, but the attribute *Italian* represent information that does not facilitate the establishment of the bridging relation. Likewise, the extra expressive load associated with the presence of an evaluative adjective seems to impede the anaphoric linkage to a certain extent.

4. Conclusion

Referential expressions in a bridging relation rely on inferential effort where a semantic or conceptual link is established between the anchor expression (e.g. *the interview*) and the indirect anaphoric expression (*the journalist*). At least for the scenarios we tested, which all used an event or product and a salient individual for the bridging relation, definite DPs prefer the unmodified version and the ratings are negatively affected by the presence of neutral and evaluative adjectives. Demonstrative DPs in turn benefit slightly from the use of a modifying evaluative adjective. This suggests that the different referential expressions serve distinct functions. Definite DPs license an anaphoric relation with a discourse-given anchor expression that relies on lexical and encyclopaedic knowledge. Demonstrative DPs signal a recognitional use that requires access to shared knowledge representations as proposed by Himmelmann (1996, 1997).

The findings indicate that form-function mapping is affected by the type of determiner used but also by the nature of the descriptive content. In our experimental stimuli, the head noun licenses a conceptual link between an event and a participant, while the demonstrative determiner activates individual knowledge in long-term memory to access a specific referent that is shared by the interlocutors. Examples for the use of demonstrative DPs in bridging contexts often include pointing gestures or prosody (Schwarz 2000: 123). This confirms the claim that the demonstrative serves the function of pointing the addressee to the relevant knowledge source, which can be enhanced by gesture or prosody. Furthermore, so-called complex anaphors such as *this fact* typically use a demonstrative determiner to signal the need for extra computational work required for establishing a propositional anaphor. Along these lines, the demonstrative in our examples is essentially used for recognitional use and to draw the addressee's attention to shared, individual knowledge.

Funding

The research for this contribution has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — Project-ID 281511265 — SFB 1252 "Prominence in Language" in the project C04 and C07 at the University of Cologne, Department of German Language and Literature I, Linguistics.

Acknowledgements

We would like to thank the two anonymous reviewers for their constructive comments that helped to improve the paper, and the editors of this volume for all their work and support. We would also like to thank Magdalena Repp and Clarissa Selegrad for their support in carrying out the rating tasks.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following glosses are used:

ANAPH anaphora
GER gerund
PAST2 past tense 2.

References

Apothéloz, Denis, and Marie-José Reichler-Béguelin. 1999. "Interpretations and Functions of Demonstrative NPs in Indirect Anaphora." *Journal of Linguistics* 24: 363–397.

Auer, Peter. 1984. "Referential Problems in Conversation." *Journal of Pragmatics* 8: 627–648.

Baayen, Harald, Douglas Bates, Reinhold Kliegl, and Shravan Vasishth. 2015. *RePsychLing: Data Sets from Psychology and Linguistics Experiments*. https://github.com/dmbates/RePsychLing

Becker, Laura. 2021. *Articles in the World's Languages*. Berlin, Boston: De Gruyter Mouton.

Burkhardt, Petra. 2006. "Inferential Bridging Relations Reveal Distinct Neural Mechanisms: Evidence from Event-related Brain Potentials." *Brain and Language* 98 (2): 159–168.

Chen, Rong. 1990. "English Demonstratives: A Case of Semantic Expansion." *Language Sciences* 12 (2–3): 139–153.

Clark, Herbert H. 1975. "Bridging." In *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, ed. by Bonnie L. Nash-Webber, and Roger Schank, 188–193. Cambridge, MA: Yale University Mathematical Society Sciences Board.

Diessel, Holger. 1999. *Demonstratives: Form, Function and Grammaticalization*. Amsterdam: John Benjamins.

Doran, Ryan B. 2012. Definiteness, Specificity, and the Multiple Uses of Demonstrative Expressions. PhD thesis, Northwestern University.

Doran, Ryan B., and Gregory Ward. 2019. "A Taxonomy of Uses of Demonstratives." In *The Oxford Handbook of Reference*, ed. by Jeanette Gundel, and Barbara Abbott, 236–259. Oxford: Oxford University Press.

Dryer, Matthew S. 2013. "Order of Demonstrative and Noun." In *The World Atlas of Language Structures Online*, v2020.4, ed. by Matthew S. Dryer, and Martin Haspelmath. Zenodo. (available online at http://wals.info/chapter/88).

Fraurud, Kari. 1990. "Definiteness and the Processing of NPs in Natural Discourse." *Journal of Semantics* 7: 395–433.

Goddard, Cliff. 1983. *A Grammar of Yankunytjatjara*. Alice Springs: Institute for Aboriginal Development.

Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski. 1993. "Cognitive Status and the Form of Referring Expressions in Discourse." *Language* 69 (2): 274–307.

Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski. 2000. "Cognitive Status and the Form of Indirect Anaphors." *Verbum* 22 (1): 79–102.

Harris, Jesse A., and Christopher Potts. 2009. "Perspective-shifting with Appositives and Expressives." *Linguistics and Philosophy* 32 (6): 523–552.

Hawkins, John A. 1978. *Definiteness and Indefiniteness. A Study in Reference and Grammaticality Prediction*. Atlantic Highlands, N.J.: Humanities Press.

Heath, Jeffrey. 1980. "Nunggubuyu Deixis, Anaphora, and Culture." In *Papers from the Parasession on Pronouns and Anaphora*, ed. by Jody Kreiman, and Almerindo E. Ojeda, 151–165. Chicago: University of Chicago.

von Heusinger, Klaus. 2011. "Specificity, Referentiality and Discourse Prominence: German Indefinite Demonstratives." *Proceedings of Sinn und Bedeutung* 15, 9–30. (Available online at https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/378.)

Himmelmann, Nikolaus P. 1996. "Demonstratives in Narrative discourse: A Taxonomy of Universal Uses." In *Studies in Anaphora*, ed. by Barbara Fox, 205–254. Amsterdam: John Benjamins.

Himmelmann, Nikolaus P. 1997. *Deiktikon, Artikel, Nominalphrase. Zur Emergenz syntaktischer Struktur*. Tübingen: Niemeyer.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36: 161–195.

Hinterwimmer, Stefan. 2019. "How to Point at Discourse Referents: On Anaphoric Uses of Complex Demonstratives." *Proceedings of Sinn und Bedeutung* 23, 487–505.

Khachaturyan, Maria. 2019. "Knowing is Belonging: Recognitional Deixis and Emergence of Common Ground in Religious Conversion." *Signs and Society* 7 (2): 186–216.

Lakoff, George. 1974. "Remarks on *this* and *that*." In *Papers from the 10th Regional Meeting of the Chicago Linguistic Society*, ed. by Michael W. La Galy, Robert A. Fox, and Anthony Bruck, 345–356. Chicago: University of Chicago.

Levinson, Steven C. 2004. "Deixis and Pragmatics." In *The Handbook of Pragmatics*, ed by Laurence Horn, and Gregory Ward, 97–121. Oxford: Blackwell.

Maes, Alfons, Emiel Krahmer, and David Peeters. 2022. "Understanding Demonstrative Reference in Text: A New Taxonomy Based on a New Corpus." *Language and Cognition* 14 (2): 185–207.

de Mulder, Walter, and Anne Carlier. 2011. "The Grammaticalization of Definite Articles." In *The Oxford Handbook of Grammaticalization*, ed. by Heiko Narrog, and Bernd Heine, 522–534. Oxford: Oxford University Press.

Nunberg, Geoffrey. 1993. "Indexicality and Deixis." *Linguistics and Philosophy* 16: 1–43.

Poesio, Massimo, and Renata Vieira. 1998. "A Corpus-based Investigation of Definite Description Use." *Computational Linguistics* 24 (2): 183–216.

Prince, Ellen F. 1981a. "On the Inferencing of Indefinite-*this* NPs." In *Elements of Discourse Understanding*, ed. by Bonnie Webber, Aravind K. Joshi, and Ivan Sag, 231–250. Cambridge: Cambridge University Press.

Prince, Ellen F. 1981b. "Toward a Taxonomy of Given-new Information." In *Radical Pragmatics*, ed. by Peter Cole, 223–255. New York: Academic Press.

Schwarz, Monika. 2000. *Indirekte Anaphern in Texten: Studien zur domänengebundenen Referenz und Kohärenz im Deutschen*. Tübingen: Niemeyer.

Skilton, Amalia, and David Peeters. 2021. "Cross-linguistic Differences in Demonstrative Systems: Comparing Spatial and Non-spatial Influences on Demonstrative Use in Ticuna and Dutch." *Journal of Pragmatics* 180: 248–265.

Wilkins, David P. 1989. *Mparntwe Arrernte (Aranda): Studies in the Structure and Semantics of Grammar*. Canberra: The Australian National University.

Wolter, Lynsey. 2006. "Bridging Demonstratives at the Semantics-Pragmatics Interface." Presentation, LSA Annual Meeting, Albuquerque, 5–8 January 2006.

CHAPTER 26

Hierarchical discourse structure in spoken narratives

Isabel Compes, Jakob Egetenmeyer & Katja Hannß
University of Cologne

Under-described languages may present diverse strategies of structuring narrative discourse. Investigations into the structure of narratives of such languages, in particular ones with a predominantly oral tradition, are still scarce, and even more so studies that relate the structure with the corresponding linguistic devices. A critical component of the hierarchical structure of discourse is captured in semantic-pragmatic relationships within and across fragments, i.e. strings of text of the length of one or more paragraphs. We investigate texts from two less described languages (Samoan, Austronesian and Chipaya, Uru-Chipayan) to delineate their hierarchical structure on a local and a global level applying three different semantic-pragmatic approaches (among them SDRT) in a comparative way.

Keywords: spoken narratives, rhetorical relations, hierarchical structure, paragraph, under-described languages, Samoan, Chipaya

1. Introduction

In this paper, we analyze the relationships between discourse units in longer strings of spoken narratives. We combine a more holistic view on discourse structure with details of local structuring. Our data is from two under-described languages, Samoan and Chipaya. Combining several approaches for the analysis of minor non-Indo-European languages is without precedents, and it allows us to reveal coarse- as well as fine-grained differences in the ways the narratives are structured.

So far, most research on discourse, its structure and the relationships across and within larger discourse units has concentrated on (major) Indo-European languages. Discourse structuring in non-Indo-European languages, especially less described ones and/or ones with low numbers of speakers, has received little to no attention at all. However, enriching the knowledge on discourse structure of

https://doi.org/10.1075/slcs.240.26com
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

oral narratives in these languages has an impact on language description practices and linguistic typology for several fundamental reasons. First, narratives are still one of the major genres collected in language documentation. Language documentation (Himmelmann 1998; McDonnell et al. 2018), in turn, lays the empirical foundation for advances in language description and linguistic typology. Thus, a better understanding of the presumably universal communicative practice of narrating and its cross-cultural variation, also in terms of discourse structure, is very much desired. Second, the Samoan and Chipaya data we use is taken from corpora of spoken narratives. This entails a specific perspective on the data: since the narratives develop in a continuous flow of speech, one of the first tasks consists in the segmentation of the discourse and the determination of units which are manageable for further analytic steps. Identifying the boundaries of units poses problems at all linguistic levels — words, phrases, clauses, sentences, and paragraphs (Himmelmann 2018; Himmelmann et al. 2018). In this paper, we concentrate on larger units, i.e. the paragraph, also called episode in narratives. Our particular focus is on how these larger units are hierarchically structured — internally as well as externally in terms of their integration into higher level structures.

We present analyses of the data based on three semantic-pragmatic approaches all dealing with different, though partly overlapping properties of discourse structuring. The first one is based on Segmented Discourse Representation Theory (SDRT, Asher & Lascarides 2003), which concentrates on local, mostly paragraph-internal, structuring, that is, the semantic-pragmatic relationships between propositions. In line with Asher and Lascarides (2003, see also Jasinskaja & Karagjosova 2020), we will refer to these relationships as "rhetorical relations" (henceforth RRs). The other two approaches deal with higher level, overarching discourse structure exploring how paragraphs are delimited and at the same time externally integrated into a global textual whole: patterns of repetition (Grimes 1972) and salience and discourse transitions (Redeker 2006; Chiarcos 2011). The latter two were already used in previous studies on Samoan and Chipaya, respectively, in which the semantic-pragmatic analyses are correlated with explicit linguistic cues (see Compes 2000, 2002; Hannß 2021). In discourse research, a basic assumption is that speakers provide cues for hearers to interpret their message, including the indication of discourse structure with its units and boundaries. Such cues belong to all linguistic levels typically forming bundles and patterns. On the other hand, the (re-)construction of the discourse as produced by speakers proceeds in an interpretative and heuristic method which is based on a semantic-pragmatic analysis, just as much as it is supported and double-checked by the analysis of linguistics cues (Serzisko 1992, 1995). Thus, while we hint at explicit linguistic cues in the sections on Samoan and Chipaya, one of our major objectives in this paper is to render the description of general narrative discourse strate-

gies more precise and, thus, to make the semantic-pragmatic arguments behind the analysis of linguistic cues transparent. Together, the three approaches allow us to describe the discourse structure of spoken narratives in a detailed manner involving the disparate but crucial components of local and global hierarchical structuring.

The paper is organized as follows. In Section 2, we give the basic theoretical background on our unit of analysis — the proposition — and then introduce the three approaches. Sections 3 and 4 analyse and discuss the Samoan and Chipaya data, respectively. Section 5 sums up our results.

2. Theoretical assumptions and introduction to applied approaches

The central aim of the paper is to arrive at a better understanding of discourse structuring within and across longer strings of text in natural language. This involves the semantic-pragmatic relationships between units in linear and hierarchical respects and the analysis of their contribution to the overall structure. An underlying challenge in natural data is determining relevant units (see e.g. Himmelmann 2006). Approaches like Segmented Discourse Representation Theory (SDRT) place the focus on semantics and investigate propositions in sequence whereas in written language, propositions generally correspond to clauses and are separated by punctuation. Here, we deal with spoken data which unfolds in a continuous flow of speech and which lacks such clear boundary signals. Our heuristic procedure in this paper is to follow SDRT and to focus on the proposition and its linguistic correlate, the predication.[1] That is, we take the content expressed in the verb complex and its arguments, which basically corresponds to a proposition, as the basic unit of our semantic analysis, assuming that this unit is overtly indicated by lexical and morphosyntactic cues. In narrative discourse, larger segments comprising several propositions are also semantic-pragmatically relevant units which we aim to determine in our analyses. In Section 2.2, we define such larger units as episodes, based on the concept of referential movement (Klein & von Stutterheim 1987). Concerning the structural correlate of episodes, we follow

1. We consider a predication to be a discourse-functional linguistic operation, "which allows a proposition to assume a self-contained linguistic form […] and enables us to report events" (see Sasse 1993: 651). According to Longacre (2007: 372), "[i]n discourse […] simple predications combine into larger units. Clauses [are] the surface structure units which correspond most closely to individual predications …". Furthermore, we conceive the content of a proposition, i.e. a state of affairs, to be centered on the verb with its arguments whose number and role is defined by the valency of the verb (see Tesnière 1959).

Longacre (1979: 115) who introduces the term *paragraph* as a larger structural unit between the local level of the predication and the global level of the whole discourse (see also Serzisko 1995).

However, our basic units of analysis, i.e. propositions, do not necessarily correspond to the well-established segmentation units of spoken language which are defined either on prosodic grounds (i.e. intonation unit and intonation groups, Ladd 1986; Chafe 1987; Himmelmann et al. 2018) or on the basis of breaks in the speech signal (i.e. pause unit, Serzisko 1992).[2] In the present paper, we use pre-annotated corpora. In reproducing the data, we stick to the predetermined units for practical reasons. These are not necessarily based on prosody, and we will shortly explain them for each corpus in the respective Sections 3 and 4.

In the following, we present a simplified version of SDRT (Asher & Lascarides 2003, Section 2.1) which we employ to account for the local relationships between propositions. Whenever helpful for classification, additional relations from Rhetorical Structure Theory (RST, Mann & Thompson 1988) are used. The other two approaches reveal patterns on an overarching global level and, thus, can show groupings of propositions that form a paragraph: patterns of repetition in an overlay (Grimes 1972, Section 2.2.1) and salience relations at discourse transitions (Redeker 2006; Chiarcos 2011, Section 2.2.2).

2.1 Local relationships between propositions according to SDRT

Verbal interaction conveys bits of information, and in addition, it conveys how these bits are related to one another. This is not only relevant on the intra, but also on the interpropositional level (Kehler 2002: 2). We may, thus, assume that in a coherent text, be it written or spoken, adjacent clauses display specific and retrievable semantic-pragmatic relationships among each other (Asher & Lascarides 2003: 1). The relationships are expected to be generally understandable given a context that is not overly determined by common ground. Furthermore, "interlocutors expect coherence as a default assumption, so that they permanently try to make sense of discourse even if connections are not immediately obvious" (Renkema & Schubert 2018: 132 with reference to Bublitz 2006: 364–365).

Examples (1) and (2) are taken from Asher and Lascarides (2003). Addressees are expected to grasp the relationships easily. In (1) the two described situations are assumed to be cotemporal. They show a so-called PARALLEL relation. In the

2. The correlation of such prosodic units with semantic-structural units (including the correlation of intonation group with the higher level unit "paragraph") is a further research question which we will not explore in this paper.

often-cited Example (2), by contrast, the second clause presents the cause of the first. This is the case of an EXPLANATION relation.

(1) *Jane speaks French. Billie speaks German.*
(Asher & Lascarides 2003: 168, adapted)[3]

(2) *Max fell. John pushed him.* (Asher & Lascarides 2003: 159)

There are many different theoretical accounts addressing the semantic-pragmatic relationships between clauses (see, for instance, Sanders et al. 2018 for an attempt to unify the annotation schemes of different frameworks; see also Zeevat 2011 on combining approaches). They diverge in terms of their background and interests, and also in their terminology. Among the more frequently used terms for the central phenomenon at hand we find "rhetorical relations" (Asher & Lascarides 2003), "coherence relations" (Hobbs 1979; Kehler 2002), "discourse relations" (Asher & Vieu 2001; Jasinskaja 2007), and in Rhetorical Structure Theory, often simply "relations" (Mann & Thompson 1988). Traditionally, accounts in the field had a special interest in anaphoric relationships of the nominal domain (Hobbs 1979). Another factor of interest is the types and numbers of relationships that need to be posited (Mann & Thompson 1988; Mann & Taboada 2005–2025). A third point is the classification of relationships to explain the availability of certain structures like VP-ellipsis and extraction in coordination (Kehler 2002). Much of the more recent research is restricted to local processes.[4] However, there have also been studies related to typology, which take semantic-pragmatic relations between sentences as a factor to analyse longer discourse spans (Longacre 1979). More recently, Egetenmeyer (in prep. a) suggests to account for episode boundaries with reference to such relations.

In this paper, we address hierarchical structure based on SDRT, which distinguishes between coordinating and subordinating rhetorical relations among propositions (Asher & Vieu 2001, 2005; Jasinskaja & Karagjosova 2020). This principle classifies rhetorical relations according to how they attach to the proposition that precedes them. SDRT is a theory rooted in formal logic. However, introducing the formalism would exceed the scope of this paper. The conception of hierarchies can be visualized in a simplified manner (see Figure 1 below) by taking the propositions as boxes, which may be connected to a previous box (the

3. For more balance across the examples shown, we changed the gender of the event participants.

4. In addition, much corpus-based and experimental research has been conducted concerning the question of how and if rhetorical relations are explicitly signaled in discourse. Since we are dealing with databases of less described languages not annotated for rhetorical relations, the discussion of this research is beyond the scope of the present paper.

preceding proposition) by one line only, either in vertical or horizontal orientation (see Asher & Vieu 2005: 595). The horizontal attachment of a proposition to a preceding one would be the case of a coordinating relation, in which case two propositions hold true on the same rhetorical plane. An example is the PARALLEL relation as in (1) (see Asher & Vieu 2005: 596). Further relations classified as coordinating are, for instance, NARRATION, RESULT and CONTRAST (see Asher & Vieu 2005: 596). By contrast, downward vertical attachment, that is, towards a subordinate plane, represents a subordinating relation. This is the case of the EXPLANATION relation as in (2) (see Asher & Vieu 2005: 596). Other important subordinating relations are ELABORATION and BACKGROUND (see Asher & Vieu 2005: 596, also for the classification of further relations). To account for passages longer than two adjacent propositions, as exemplified by (3) below, two further descriptive components are needed. The first one concerns a restriction on possible attachment sites. It has become known as the "right-frontier-constraint" (Polanyi 1988).[5] On the basis of a hierarchical understanding of a developing discourse, she argues that different planes may be co-present at a certain point in discourse (Polanyi 1988: 602–603). She states that all but only the constituents bordering on the right are available for subsequent attachment (see Polanyi 1988: 613, 615). This means that due to the hierarchical diversification of propositions, content may be added at different discursive planes. The prerequisite is rhetorical subordination at a previous point in the discourse. More specifically, in the case of a subordinating relation, two propositions (or visually, two boxes) are given at the right frontier. If in the discourse, a subordinating relation is followed by another subordinating relation, three attachment sites would be available. Secondly, a very helpful term is the superordinating relation (see Egetenmeyer in prep. b). It refers to the constellation where the discourse returns to a superior plane, which was introduced into the structure at an earlier point of the discourse. Again, the prerequisite is rhetorical subordination at a previous point in the discourse. The discourse can continue on this subordinate plane maintaining the availability of the superior plane, which the structure can then return to.

The following classic example by Asher and Vieu (2001: 1) is frequently cited to make the basic functioning of rhetorical relations more readily accessible (see, for instance, Asher & Lascarides 2003: 8; Asher & Vieu 2005: 592; Becker & Egetenmeyer 2018: 45–46).

5. As a reviewer reminds us, the view of SDRT on discourse, which involves the right frontier constraint as a core principle, has been challenged by Wolf and Gibson (2005) for naturalistic data.

(3) a. *John had a great evening last night.*
b. *He had a great meal.*
c. *He ate salmon.*
d. *He devoured lots of cheese.*
e. *He then won a dancing competition.*
f. ?*Then he had a great dessert.*
f′. #*It was a beautiful pink.*
f″. *John had lots of fun.*

Using the above-mentioned boxes for propositions, the rhetorical structure of clauses (3a)–(3e) can be visualized as in Figure 1 (see Asher & Lascarides 2003: 8 and Becker & Egetenmeyer 2018: 46, also for the following description). Clause (a) is elaborated by (b) to (e). Clause (b) is itself elaborated by (c) and (d), which are related among each other by a NARRATION relation. At the point in the text when (d) has been uttered, three propositions border with the current right frontier on three different hierarchical planes: (a), as indicated by the dotted line within the respective box, (b) and (d). If John had had an expresso following the cheese, the attachment would be realized on the lowest plane. However, his evening continues with something other than nutrition, namely a dancing competition. Therefore, a superior discourse plane is continued, as (e) is attached to (b) by a NARRATION relation. This is a superordinating relation as the attachment is to a proposition higher in the structure than the directly preceding proposition (d). Thus, the structure returns to the (already given) superior plane. After (e) has been uttered, only (a) and (e) are located at the right frontier. Consequently, Asher and Vieu (2005: 592–593) discard the continuation of (f) as not fully natural,[6] since the dessert would be eaten as part of the meal and not after a lengthy different evening event like a dancing competition. Furthermore, (f′) is not a possible continuation if the pronoun *it* is intended to be coreferential with *salmon* (see Asher & Vieu 2005: 593). This is a direct consequence of the right frontier constraint, since (c) is not available for attachment at this point in the discourse structure. Finally, (f″) is a possible continuation of the text. As Asher and Vieu (2005: 593–594) point out, depending on the intended reading, it may be attached either to (a) or to (e).

One of the simplifications of our analysis is how we account for the forward-oriented BACKGROUND relation (see Sections 3.2 and 4.2). Asher et al. (2007) implement it as a framing topic (FT) relation, to which the foreground is attached

6. This view is in accordance with an assumed typical dinner script. Note that this script is culturally biased. The implied expected order of events is to have cheese followed by dessert. This order is normal in Great Britain, while in France, for instance, it is typically the other way around. Assuming the French order, attaching (f) would be even more problematic.

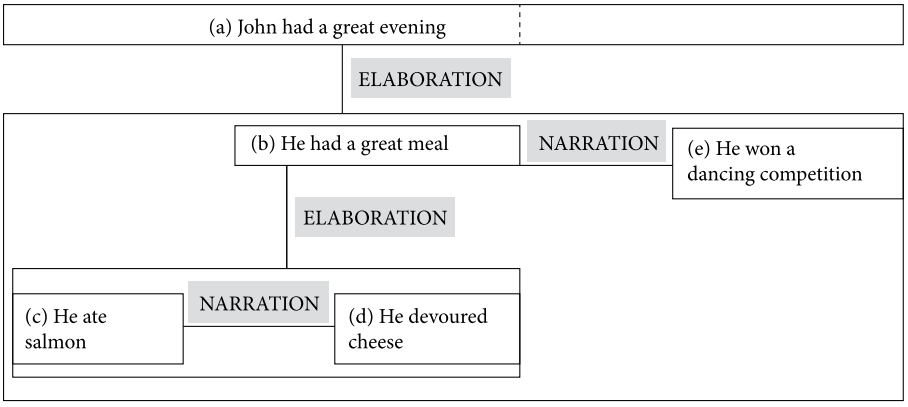


Figure 1. Visualization of rhetorical structure of (3) (Asher & Lascarides 2003: 9; Becker & Egetenmeyer 2018: 46; with small but important adaptations)

by means of an ELABORATION relation. However, this would (i) cloud the impression of the hierarchical relations accounted for here, and (ii), make textual development (see Egetenmeyer 2024) less evident in the analysis. Instead, in our analysis we maintain the textual order and the propositional hierarchy of the forward-looking BACKGROUND relation. We do so by implementing a framing topic (FT), that is an empty element (so to speak, an empty box) to which the BACKGROUND is anchored (downward branching) and from which the non-backgrounded storyline continues. This second relation from the framing topic to the following proposition will be marked here as FOREGROUND to underline the interdependency. In the quantification of rhetorical relations, only the downward branching (marked as BACKGROUND) will be counted and not the rightward branching from the (empty) framing topic to the following element.

2.2 Semantic-pragmatic approaches with a holistic view

In this section, we will briefly introduce more holistic semantic-pragmatic approaches that played a supporting role in the previous studies in which a major objective was the determination of larger discourse units. In general, these approaches belong to traditions of discourse analysis which, on the one hand, have a strong structural interest (Halliday 1967a, 1967b, 1968; Longacre 1976, 1983; Grimes 1975; Givón 1983), and which, on the other hand, focus on narrative storytelling (Labov 1972, see also Bamberg 1994). Both perspectives on discourse structure have proven fruitful for typology-informed research. Here, we want to highlight the notion of referential movement introduced by Klein and von Stutterheim (1987) which is used to identify boundaries in discourse (see also the

similar notion of topic continuity, Givón 1983). According to Klein and von Stutterheim, referential chains of participant, time and place constitute a narrative text world. Changes in reference, i.e. the referential movement, reveal continuities and discontinuities in the discourse providing the semantic-pragmatic evidence for larger discourse units. These units are called episodes or scenes and correspond to the *paragraph* as a mid-level structural unit (see Longacre 1979 and Serzisko 1995 in Section 2). Across episodes, the parameters of participant (and place) change. Within an episode, the parameters of participant and place remain constant (see anaphoric relations mentioned in Section 2.1). The parameter of time, however, is intrinsically dynamic. Each of the events reported in an episode (in the unmarked order of events) refers to a specific time interval *t2* which is chronologically subsequent to the time interval *t1* of the event reported in the preceding utterance (Klein & von Stutterheim 1987: 168). We assume that this conception of textual reference in narrative is uncontroversial and that it constitutes the basis of all narrative analysis while, at the same time, intricacies of this simplified picture will be illustrated by the text examples in Sections 3 and 4.

2.2.1 *Patterns of repetition (Grimes 1972)*

The analysis of the selected Samoan narrative (see Section 3) benefits from concepts presented in Grimes (1972). Grimes considers lexical information and patterns of repetition to analyse the discourse structure of texts from several less-described languages (see also Fleischman 1986 on Old French). In discourse analysis, repetitions are examined as ties creating both cohesion and coherence (see Tannen 1987; Johnstone 1994). For one, repetitive patterns are determined on the basis of the repetition of lexemes where the exact word iteration is a copy of content but is also a formal indicator of cohesion. Secondly, variation and paraphrase are analysed as repetitions. These are identified on the basis of semantic criteria which create coherence. Grimes adopts a global perspective and adds a new dimension to discourse insofar as he tracks repetitions over longer distances and insofar as he interprets them as structural features of the discourse. In this way, he establishes a unique discourse structure — the *overlay* — which is typical for certain oral narrative traditions and which stands in opposition to the *outline*. We shortly present the main features of this structure but start with the *outline* which has been studied more frequently and is typical for Indo-European languages (with a writing tradition).

An outline corresponds to an episode as conceived by Klein and von Stutterheim (1987), i.e. the time parameter is intrinsically dynamic, and participant and place remain constant. Furthermore, in an outline speakers organize their discourse along two dimensions of textual linking: a coordinating and a subordinating one which are further subclassified in relations such as temporal sequence, alter-

native, description, explanation (Grimes 1972: 513–516). Thus, Grimes' outline is an early approach to rhetorical relations and corresponds closely to the conceptualization of hierarchical relations in SDRT (see Section 2.1).

In an overlay, speakers work with a third dimension that results from repetition and coreference: Strings of text are repeated, paraphrased and expanded. Grimes refers to these strings as *planes*. To distinguish the concept from the notion of planes in SDRT (see Section 2.1), we will refer to them as *repetition planes* in this paper. Importantly, no such string of text "gives all the information contained in the whole overlay; one [repetition] plane is not just a summary of another" (Grimes 1972: 516). This means that, on a global level the repeated information links the repetition planes and integrates them into a structural whole — the overlay.

In the repetition planes of an overlay, the three parameters of participant, place and time are implemented in a specific way (Grimes 1972: 519–520). Basically, all three parameters are established in the first repetition plane, and throughout the overlay, the parameters of participant and place conform to the usual behaviour of the referential movement. However, the parameter of time exhibits the following particular characteristics:

i. Whenever a new repetition plane starts within an overlay, the reported time interval of the narration jumps back to the time interval of the initial plane, thus interrupting the linear temporal progression of narrative time (= interrupted intrinsic dynamicity, temporal loop).
ii. A repetition plane is internally structured according to the principles of an outline, i.e. the temporal order of the reported events proceeds sequentially forward (= intrinsic dynamicity).

Coming back to the metaphor of dimensions, an overlay resembles a multi-storey building with its three dimensions where each floor is laid over the preceding one and in each floor some structures of the lower floor are repeated. These repetitions are the load-bearing walls which support the entire building while the make-up of each floor internally allows for variation.

2.2.2 *Salience and discourse transitions (Redeker 2006, Chiarcos 2011)*

A previous study on discourse coherence in Chipaya investigates the Chipaya salience-marking enclitics *=l*, *=m* and *=z* (henceforth SMEs; see also Section 4) as "discourse operators" (Redeker 2006: 341) against the theoretical background of "salience" and "discourse transitions" (Hannß 2021).

According to Chiarcos et al. (2011: 2), "[s]alience defines the degree of relative prominence of a unit of information, at a specific point in time, in comparison to the other units of information" (on prominence, see also Himmelmann & Primus

2015). One of two basic types of salience is "*speaker salience* (*importance/newsworthiness*). Speaker salient information is speaker-private and relevant, e.g. new for the hearer, not predictable or something the speaker wants to put special emphasis on" (Chiarcos 2011: 107; emphasis in the original).[7] The SMEs attach to elements that are new to the hearer, important and/or not predictable and are thus markers of speaker salience. As "salience is a necessary condition for shifts of attention" (Chiarcos 2011: 109), the SMEs also indicate shifts in salience, and they do so at discourse transitions (see Hannß 2021). There are two types of discourse transitions: "*paratactic transitions* between segments that follow each other at the same level" and "*hypotactic transitions* involving interruption or suspension of an incomplete unit with parenthetical material" (Redeker 2006: 344–345; emphasis in the original). Next-segment and/or end-of-segment markers indicate paratactic transitions, while parenthetical material that causes hypotactic transitions includes, among others, quotations (Redeker 2006: 344–345).

Discourse transitions are, of course, boundaries in a discourse, which are characterized by referential movement (see Klein & von Stutterheim 1987; see also Section 2.2), i.e. the attention shifts from one referent to another — and often newly introduced — referent. Thus, by indicating shifts in salience, the SMEs also point to a change in reference, thereby evidencing paragraph boundaries.

3. Samoan

Samoan (ISO 639-3: smo, Glottocode: samo1305, Polynesian) is the native language of around 430,000 speakers worldwide (Eberhard et al. 2020). It is a verb-initial isolating language with syntactic ergativity. Linguistic descriptions include dictionaries and grammars (see among others Milner 1993; Mosel & Hovdhaugen 1992). The Samoan text *'O lo'u ōlagamai* is part of a larger corpus of spoken narratives which are segmented into pause units (=PU).[8] We selected two passages (PUs 07–24 and 73–87, length of pauses in brackets, numbering corresponds to the PUs of the whole text). For reasons of space, we only provide the English translation for the introductory and framing PUs and for some less relevant PUs. Note that PUs may be smaller or larger than units which correspond to proposi-

7. The counterpart, which we will not focus on, is "hearer salience" (Chiarcos 2011: 108).

8. The texts were tape-recorded by Even Hovdhaugen and Ulrike Mosel during research stays in Samoa between 1982 and 1990 (cf. Mosel & Hovdhaugen 1992) and processed by Isabel Compes (see details in Compes 2000). Pauses are relatively easy to determine, and their relevance for discourse structure is undisputed (see Ladd 1986; Chafe 1987; Serzisko 1992). The guideline for a perceptible pause was 0.2 sec.

tions, and we sometimes present several PUs conveying one proposition in one line only. Furthermore, some PUs were split up to represent propositions numbered (a) and (b) respectively.

The narrative is the life story of a female speaker of 5 min. 20 sec. length. She talks about the German occupation of Samoa at the beginning of the last century with themes such as persecution and death, resistance and revolution, against the background of two contrasted groups: the group of the white invaders and the group of Samoans.

Excerpt 1.

07–08 'the things I saw (0.95) in my life (1.1)'

09–10 *le* ***o'o*** ***mai*** (0.7) *o* *le* *sāuā-ga* (2.55)
SPEC.SG **reach DIR** POSS.NAL SPEC.SG be.cruel-NR
'the reaching hither of the (0.7) act of cruelty (2.55)'

11–12 'of the Germans (0.55) to Samoa (1.4)'

13 'I saw (that) (0.9)'

14 *sā* ***ō*** ***mai*** *ia* *papālagi* *m-ā* *lakou faga* (0.9)
PST **go.PL DIR** that.PL white.man with-POSS.AL 3.PL gun
'(here) came those white men with their guns (0.9)'

15–16 *e* *kulikuli* *ai* *makai* *o* (0.25) *Sāmoa* (1.85)
GENR pursue.unfairly by.it titled.head POSS.NAL Sāmoa
'to chase with them the chiefs of (0.25) Samoa (1.85)'

17 *sā* ***ō*** *i* *solofagua* (0.8)
PST **go.PL** LD horse
'they came on horses (0.8)'

18 ***ō*** ***mai*** *i* *pō* (0.9)
go.PL DIR LD night
'came here at night (0.9)'

19–20 *kuli* *ka'amilo* (0.9) *fale* *o* *'āiga* (1.35)
chase go.around house POSS.NAL family
'chased went around (0.9) (in) the houses of the families (1.35)'

21 *e* *va'ai* *mea* *o* *makai* (0.55)
GENR look.at place POSS.NAL titled.head
'to look for the places of the chiefs (0.55)'

22–23 ***ma*** *ave* (0.85) *i* *luga* *i* (0.6)
and take LD upwards LD
'and took them (0.85) upwards to (0.6)'[9]

9. A comment on this PU in the native speaker transcription clarifies what is meant here: 'put matai on horses'.

24 *e lafo a lākou lāfoga* (2.4)
GENR throw POSS.AL 3.PL tax
'to/so that they pay their tax (2.4)'

Excerpt 2.

73–78 'well my eyes also saw the days of the epidemic (1.4) the epidemic (0.85) the troops they buried the people of Samoa (1.65) the Samoans (0.55) it was like the treatment of (1.05) of animals (1.5)'

79 ***'a ō-mai*** *gei ā le vaega'au i l-o kākou gei fale* (2.7)
when go.PL-DIR here EMPH SPEC.SG troops LD SPEC.SG-POSS.NAL 1.PL.INC this house
'when the troops came here really to our own house here (2.7)'

80 *ua feoki pē ko'a-* (1.2)
PERF die.PL or CLASS.human
'there died about [people-]... (1.2)'

81a *kolu pē ko'a-fa gi-si*
be.three or CLASS.human-be.four NSPEC.PL-other
'three or four people and some more'

81b *ua feoki i l-o kākou fale* (0.85)
PERF die.PL LD SPEC.SG-POSS.NAL 1.PL.INC house
'they died in our house (0.85)'

82a *gā gā* ***ō-mai*** *ā*
only that **go.PL-DIR** EMPH
'they just came indeed'

82b–83 *fekoso i fafo* ***ma*** *ka'eli* (0.95) *(ia) ia pū* (0.85)
pull.PL LD outside **and** dig.PL this.PL this.PL hole
'pulled them outside and dug (0.95) these holes (0.85)'

84–85 ***ma*** *lafo fa'akasi lava iai ma* (0.45) *ma fala la ma* (0.75)
and throw together EMPH to.it with with mat ? with
'and threw them all together in there with (0.45) with mats ? with ... (0.75)'

86–87 *i le fai-ga o Sāmoa* (0.7) *e vaega'au* (1.9)
LD SPEC.SG do-NR POSS.NAL Sāmoa ERG troops
'in the treatment of the Samoans (0.7) by the army (1.9)'

3.1 Analysis according to patterns of repetition

The text constitutes a network of information characterized by iteration, paraphrase, variation and specification with abundant repetitions. The following lexical chain illustrates varying references to the Germans: *Siamagi* 'Germans' – *papālagi/palagi* 'whites, white men' – *vaega'au* 'troops' – *fitafita* 'soldiers'. In Compes (2000, 2002), the text was analyzed according to Grimes' approach (1972, see Section 2.2.1) in terms of an overlay which comprises seven repetition planes and which is complemented by one outline. For the text analysis, we select the first two repetition planes (see 14–16) and (17–24) in Excerpt (1)) and the outline (see (79–87) in Excerpt (2)), each presented with minimal preceding introductory context. In Excerpt (1), the propositions report on events of the first topic: "the occupation". In Excerpt (2), the propositions report on another topic with a series of events: "the epidemic". However, the overall theme of the invasion of whites to Samoa remains in focus throughout the whole text.

On the basis of referential movement which evinces the specific implementation of the text world parameters described for overlays in Section 2.2.1, we distinguish four paragraphs, two in each excerpt. The first paragraph of Excerpt (1) (repetition plane 1, (14–16)) starts by a shift in time from speaker's now to past time and establishes participants and place. The second paragraph (repetition plane 2 (17–24)) jumps back again to this temporal starting point, reestablishes participants and place and reports the same events as the first paragraph (see the events "come and chase") providing further details. Thus, as characteristic for overlays, the intrinsic dynamicity is interrupted and the parameter of time is in a loop while the parameters of participant and place as given by the overall theme remain constant across repetition planes.

For the outline in Excerpt (2), we note an explicit update of the time parameter (see the introduction of the second topic "days of the epidemic" in (73)) which is followed by a series of sequentially ordered events. Furthermore, the parameters of participant and place do not change but they are re-established twice in propositions which do not report events on the storyline but repeat the overall theme of the invasion (see (79) and (82a)). We interpret this explicit reestablishing of participant and place as a marking of paragraph boundaries.

Two linguistic cues underpin this semantic-pragmatic analysis relying on lexical material. For one, this is the use of an unspecific motion verb with the directional particle *mai* 'hither; towards the DC', and secondly, these are the connectors *'a* 'when' and *ma'* and'. The construction [unspecific motion verb + DIR] is the primary linguistic cue found at paragraph boundaries. In the introduction of topic 1 (09–10), the unspecific motion verb *o'o* 'reach' is used, but in most cases, it is the verb *ō* 'go.PL'. It marks the first repetition plane (14), the second one

(17–18), and it cooccurs with the division of the outline into two paragraphs (79) and (82a). Thus, the use of this construction confirms the segmentation into four paragraphs.

The text structure is closely connected to the text world parameter of place. While the motion verbs constitute lexical repetitions, they are modified by the deictic-directional particle *mai* 'hither'. This deictic particle plays an important role for the text world parameter of participant. In the introduction (09–10), source and goal of the movement are specified — the Germans to Samoa — where Samoa is the deictic centre. This means that, in the use of the deictic-directional particle *mai* 'hither', the parameter of place is interconnected with that of participant. This constellation of place and participant is kept throughout the whole narrative. The particle *mai* 'hither' reestablishes both place and participant at every boundary of a paragraph (see (14), (18), (79), (82a)). In contrast, the agent of the movement (the Germans or paraphrases like 'the whites') is not always explicitly mentioned in an NP, nor are personal pronouns used (see (17–18), (82a)). Instead, the deictic-directional particle ensures the anaphoric chaining across subepisodes by mapping the relevant participant variable (of the Germans) on the place variable (*here*, the location of the speaker), and in this way functions as an anaphoric marker.

It is striking that almost no connectors are used in the excerpts. An exception is found in (79). The connector *'a* 'when' is an explicit boundary marker of a paragraph providing the background at the beginning of the outline on a global level. In addition, the internal cohesion of paragraphs is reinforced by the connector *ma* 'and' which is used three times and functions as coordinator on a local level, connecting adjacent propositions (see (22) in repetition plane 2, and (82b) and (84) in the outline).

3.2 Analysis according to SDRT

We summarize our analysis according to SDRT in Figure 2 for the two repetition planes and in Figure 3 for the outline.[10] As seen in Figure 2, the whole configuration displays a vertical structure. The two repetition planes are linked to the introductory text by a subordinating relation (ELABORATION). What is striking is that also within the two repetition planes most relations are subordinating (ELABORATION and EXPLANATION) while only two coordinating relations (NARRATION) are found. The propositions (18) to (19–20) and (21) to (22–23) stand in a superordinating relation.

10. For the formalism of these figures as well as the one in Section 4, the reader is referred to Section 2.1.

As seen in Figure 3, also the outline is linked to the preceding text by a subordinating relation (ELABORATION). Within the outline on the other hand, coordinating relations (FOREGROUND and NARRATION) are balanced with subordinating relations. Moreover, two of the propositions with a subordinating relation (79, 82a) have the function of localizing a set of following events temporally (forward-looking BACKGROUND, followed by a corresponding FOREGROUND relation, see 2.1 on FT). Thus, the outline shows a more horizontally organized lay-out.

Coming back to the repetition planes in Figure 2, an especially peculiar case is (17). It links up to (14) in an ELABORATION relation, and at the same time, it reactivates the storyline by repeating lexical material from the introduction in (13). In general, the storyline reporting sequentially ordered events is constituted by coordinating RRs as NARRATION, RESULT etc. wherein subordinating RRs as ELABORATION and EXPLANATION propositions are embedded on a lower hierarchical plane. Superordinating RRs allow to jump back to the storyline which is usually accomplished by a coordinating RR (see Section 2.1). In contrast, in this text the storyline proceeds step by step vertically downward on hierarchically lower planes in a sequence of subordinating relations. Thus, in its linking to the storyline, certain instances of ELABORATION relations have functions comparable to a superordinating relation (see (14), (79) and (86)). Furthermore, the repetitive patterns add a further dimension to the structure which are captured in the graphs by additional links (see dotted lines, compare also (13) to (79) and (82a) in Figure 3). In terms of RR, the propositions linked by repetition resemble the LIST relation.

This peculiarity coincides with the overlay structure described above in which paragraphs are embedded into a thematic whole by repetitive patterns. In the overlay structure, a sequence of ELABORATIONS on a theme is reported on the storyline, marked by extensive use of repetitions. On these hierarchically lower repetition planes, only short sequences of events in a NARRATION relation are reported. Crucially, at critical points of branching in the structure, there is no temporal advancement, but instead the narrator gives further details with respect to the described situation.

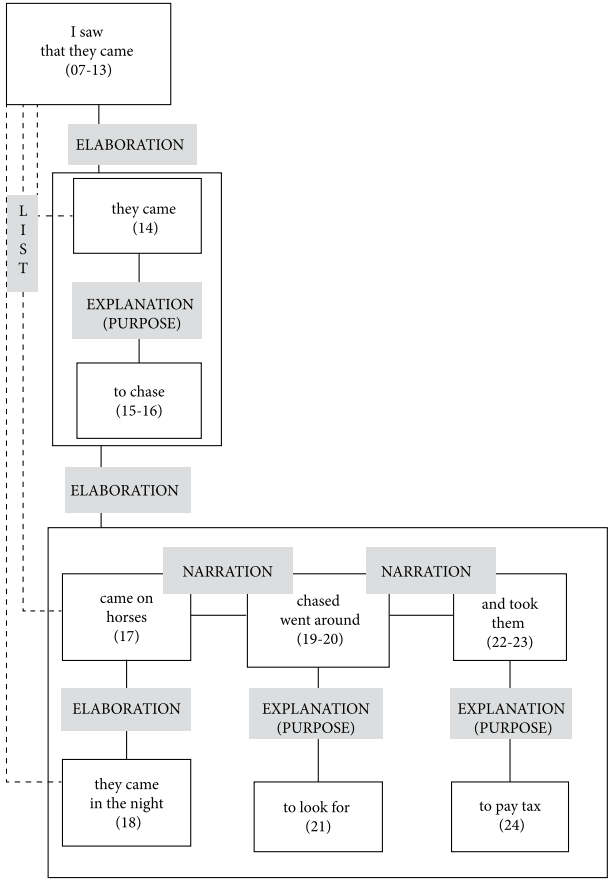


Figure 2. Analysis of the Samoan Excerpt (1) according to an SDRT approach

3.3 Comparison of approaches

Table 1 summarizes the results of our text analysis presenting the SDRT approach in the first column, the overlay structure in the second and the correlation with linguistic coding in the next two columns (verb construction and connector). It allows us to compare the contribution of the approaches to determine the hierarchical order of propositions on a local level and to identify paragraphs on a global level.

Comparing the SDRT approach in relation to the overlay, the former does not reveal the four paragraphs as given in the excerpts of the latter. Interestingly, all paragraphs start with a subordination relation: two with an ELABORATION (14 and 17), two with BACKGROUND (79 and 82a). Coordinating relations only appear internally to these units. A pattern — subordination at the boundary, coordination

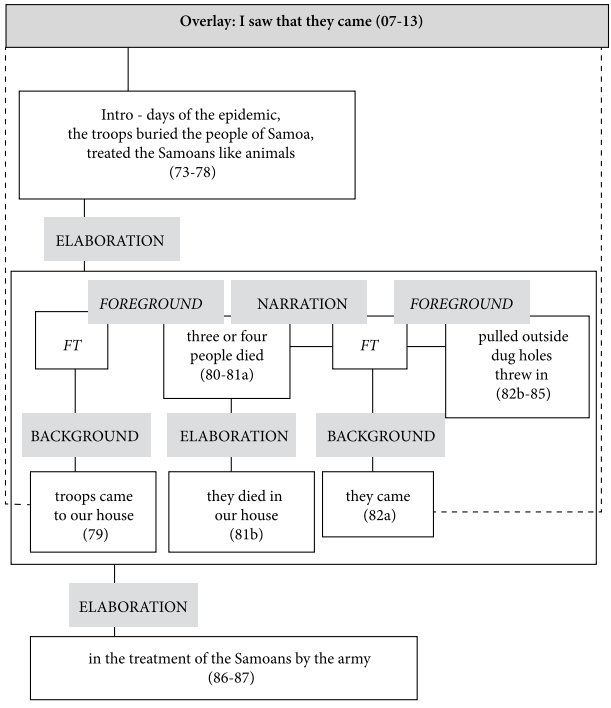


Figure 3. Analysis of the Samoan Excerpt (2) according to an SDRT approach

unit-internal — may therefore be interpreted as a semantic-pragmatic cue to paragraph structure within an overlay.

The overlay structure was already correlated with the linguistic coding in Section 3.1, in particular regarding the pattern [unspecific motion verb + DIR] at paragraph boundaries. To underline this correlation, we would like to highlight the role of the connectors which exhibit a clear functional contrast between global and local level, which, in turn, also correlates with SDRT relations. For one, *'a* 'when' appears at a paragraph boundary on a global level. The boundary correlates with the realization of a subordinating BACKGROUND relation. In contrast, *ma* 'and' consistently only occurs internal to a paragraph and correlates with the coordinating NARRATION relation in SDRT.

4. Chipaya

Chipaya [ISO 369-3: cap, Glottocode: chip1262, Uru-Chipayan] is an indigenous language of highland Bolivia, spoken by 1,000 to 2,000 people (Adelaar 2007: 19; DobeS). It forms part of the small Uru-Chipaya language family, the other two

Table 1. Comparison of SDRT, overlay approach and linguistic coding

No	Semantic-pragmatic approach		Linguistic coding	
	Local — SDRT	Global — Overlay	Verb construction	Connector
Excerpt (1) (=Repetition planes)				
7–13		Intro Topic 1	unspec. motion + *mai* 'DIR	
14	Subordination ELABORATION	1 = Repetition Plane 1 → Shift to past time, Establishing of participant and place	unspec. motion + *mai* DIR	
15–16	Subordination EXPLANATION			
17	Subordination ELABORATION	2 = Repetition Plane 2 → Time loop, Reestablishing of participant and place	unspec. motion + prop	
18	Subordination ELABORATION		unspec. motion + *mai* DIR	
19–20	Coordination NARRATION			
21	Subordination EXPLANATION			
22–23	Coordination NARRATION			*ma* 'and'
24	Subordination EXPLANATION			
Excerpt (2) (=Outline)				
73–78		Intro Topic 2 → Time update		
79	Subordination BACKGROUND	3 → Reestablishing of participant and place	unspec. motion + *mai* DIR	*'a* 'when'
80–81a	Coordination FOREGROUND			
81b	Subordination ELABORATION			
82a	Subordination BACKGROUND	4 → Reestablishing of participant and place	unspec. motion + *mai* DIR	
82b–85	Coordination FOREGROUND			
	NARRATION			*ma* 'and'
	NARRATION			*ma* 'and'
86–87	Subordination ELABORATION			

members of which have already become extinct (Hannß 2024: 265, referring to Adelaar with Muysken 2004: 622, Hannß 2008: 5). Chipaya is agglutinating with some fusional tendencies (Cerrón-Palomino 2009: 47). The text to be discussed in the following is taken from the Chipaya DobeS corpus.[11]

The SDRT and salience relations approaches are exemplified by 17 clauses from a Chipaya folk story, where each clause presents a proposition.[12] Clauses that are morphosyntactically related are numbered with (a), (b), etc. in the text below (e.g. (2a), (2b) and (2c)). In Chipaya, a clause basically consists of a predicate and its arguments.[13] This is consistent with our definition of a predication as the structural correlate of a proposition (see Section 2). Additional lexical and morphosyntactic indicators of clausehood are:

- connectors, such as *nuẓkiṣ* 'thus, so' and *neqhṣtan* 'then' in clause-initial position
- the subject (following the connectors, if present), which carries the topic marker *=ki*
- the declarative marker *=tṣa* marks a predication (Cerrón-Palomino 2006: 166–167) and attaches to the predicate of a clause. Since Chipaya is an SOV language, the declarative marker thus 'closes' a canonical declarative clause.

Note also that the Chipaya corpus was segmented into clauses upon transcription and translation by Chipaya native speakers. Although native speaker competence is not infallible, it offers an additional orientation for structuring the Chipaya text into clauses. Recall that we assume clauses to be structural correlates of propositions, our basic units of investigation (see Section 2).

The text was produced by a middle-aged male native speaker of Chipaya and is about the fox and the armadillo, both competing over the king's daughter, who is pregnant by the armadillo. In the part of the story discussed here, the king wants to find out who the father of the princess' child actually is. The excerpt is preceded by another eight clauses (not shown here), which briefly describe how fox and armadillo both desire the princess. Thus, line numbering is ours and (1) is not the first clause of the entire story. The remainder of the story (not displayed, either) describes three tests which fox and armadillo have to pass to win the princess. In this, fox and armadillo become the central characters of the story, while the king, the princess and the child are entirely backgrounded, having no appearances or hardly any at all.

---

11. See *The Language Archive* (TLA) at: https://hdl.handle.net/1839/77f5bf63-956a-4c5a-bb43-c4aeead6c50e.

12. The entire folk story consists of 89 clauses and is 10:23 minutes in length.

13. Note, however, that activated referents may be left unexpressed.

Excerpt 3.

1 ***nuẓkiṣ*** *ni wawa=ki pʰutʃ-kiṣ paq-tʃi=ki=tṣa*
thus the child=TOP belly-LOC.PROX grow-COMPL.3SG.M=REP=DECL
'Thus, the child grew in the belly.'

2a ***nuẓkiṣ*** *wawa=ẓ matʰ-tan*
thus child=**SME** be.born-SR/-COR-SIM
'Thus, when the child was born,'

2b *wawa matʰ-ta=kʰen*
child be.born-PST=REASON
'because the child was born,'

2c *ni rey=ki tuz kʰi-tʃi=ki=tṣa*
the king=TOP so say-COMPL.3SG.M=REP=DECL
'the king said so:'

3 *xe:k-ẓ=ta=t xaṣi ti wawa*
who-POSS=Q.INFORM=? now this child
'Whose child is this now?'

4a *ti wawa=x perṣuna epʰ pax-la kʰi-kan*
this child=CONJ personally father recognise-HORT say-SR/+COR+SIM
'Saying "Let this child personally recognise the father"',

4b *tʰappatʃa ti yoq-kiṣ ẓeɭ-ɲi animala-naka qʰaw-ṣi=ki=tṣa*
all this earth-LOC.PROX be-AG animal-PL call-COMPL.3SG.M=REP=DECL
'he called all the animals that exist on this world.'

4c *ni wawa epʰ perṣuna paxa-xu*
the child father personally recognise-IMP.IND
'The child should personally recognise the father.'

5a ***neqʰṣtan*** *ni wawa tʰuɭ-ẓ-ti-kiṣ ni:-ẓ yukʰ-kiṣ*
then the child seat-REFL-ASC-LOC.PROX 3SG.M-POSS face-LOC.PROX
wat-qat-tʃi=ki=tṣa
go.by-CAUS-COMPL.3SG.M=REP=DECL
'Then, where the child was seated, [the king] made [the animals] go by in front of him (i.e. the child).'

5b *ni animala wezla-naka tʰapaman wezla-naka*
the animal bird-PL all bird-PL
'The animals, birds, all birds,'

5c *qʰaṣ-kiṣ ẓeɭ-ɲi qʰoɲ-kin ẓeɭ-ɲi*
water-LOC.PROX be-AG dry-LOC.DIST be-AG
'that are in the water, that are on dry land.'

6 ***nuʐkiʂ*** *ni wawa ana ʂinta wawa=q xe:k=mi ep^{h} pax=ki=tʂa*
thus the child NEG single child=CONF who=ADD father recognise=REP=DECL
'Thus, the child did not recognise a single one as father, either.'

7 ***nuʐkiʂ*** *qiti=ki=ʐ t^{h}on-tʃi*
thus fox=REP=**SME** come-COMPL.3SG.M
'Thus, the fox came.'

8a *qiti ni waw-ʐ yuk^{h}-kiʂ aɲtʃa k^{h}urʂ x^{w}at~x^{w}at-ʐin=ki=xa*
fox the child-POSS face-LOC.PROX very tail wag~wag-BEN.V=REP=CONJ
'The fox wagged the tail wildly in front of the child,'

8b *tata=ʐ k^{h}e: tata=ʐ k^{h}e: k^{h}i:-kan*
father=**SME** say.IMP father=**SME** say.IMP say-SR/+COR+SIM
'saying "Say "father", say "father"!"'

9 *wawa=ʂte tʃʔuxu=qaʂ ana ʂinta tʂ^{h}ulu=mi tʃi:(-tʃi=ki=tʂa)*
child=CON call=RESTR NEG single what=ADD say(-COMPL.3SG.M=REP=DECL)[14]
'But the child did not call out once, either.'

10 ***neq^{h}ʂtan*** *wiri t^{h}iya=ʎa kerka=ki=ʐ t^{h}oɲ-tʃi*
then last aside=LIM armadillo=REP=**SME** come-COMPL.3SG.M
'Then, at last, the armadillo came slinking in.'

4.1 Analysis of the Chipaya text

In the following, we first present an analysis of the Chipaya text based on salience relations at discourse transitions, before turning to an SDRT approach (Section 4.2).

Based on an analysis according to principles of transitions (Redeker 2006) and salience (Chiarcos 2011), we suggest four paragraphs. The first paragraph consists of (1) only. However, we argue that it is set off from the following paragraph by its semantic function — (1) is an introduction to the following part of the discourse, summarizing the preceding events, which result in the princess' pregnancy. There is thus referential movement of the parameters of time and participant and hence a paratactic transition from (1) to (2a). The second paragraph covers (2a) to (6), in which the king and the child are in focus. We propose another paratactic transition from (6) to (7), where clauses (7) to (9) present the third paragraph. There, reference shifts from the king and the child to the major character of the fox, which appears in (7). Finally, the fourth paragraph begins at (10) (the remainder of the paragraph is not displayed here).[15] The armadillo

14. The morphology in parentheses was added when the narrator transcribed and translated the text, i.e. it was not produced in the original oral version.

is another central character, entering the stage in (10). From (9) to (10), we find once again referential movement — from the fox to the armadillo — and therefore another paratactic transition. Within these four paragraphs, there is the following substructuring. The second paragraph from (2a) to (6) is partitioned into two subparts: the first covers clauses (2a) to (4c), in which the king wonders about the child's father. The second subpart ((5a) to (6)) describes the parade of the animals.

This semantic-pragmatic structuring correlates with the distribution and co-occurrence of the SMEs and the connectors. Thus, the summary of the preceding events in (1) is introduced by *nuẓkiṣ* 'thus', which formally separates (1) from the previous as well as the following paragraphs. The referential shift to the king and the child in (2a) is formally indicated by — again — use of the connector *nuẓkiṣ* 'thus' as well as the SME *=ẓ* on *wawa* 'child'. Internal semantic-pragmatic structuring of the second paragraph, i.e. the parade of the animals ((5a) to (6)) coincides with the connector *neq^{h}ṣtan* 'then' in (5a). The second subpart and with it the paragraph containing it close with a further use of *nuẓkiṣ* 'thus' in (6). The shift from king and child to the major character of the fox, whose appearance initiates the third paragraph in (7), coincides with a connector and the SME *=ẓ* on *qiti* 'fox' as does the arrival of the armadillo in (10). This fourth and last paragraph is formally initiated by the connector *neq^{h}ṣtan* 'then', and another SME, attached to *kerka* 'armadillo'. Also, we would like to point to the "quotation" in (8b) as an instance of a hypotactic transition (Redeker 2006:344; see Section 2.2.2). It presents direct speech by the fox, which is marked twice by the SME *=ẓ* (by way of repetition), making the audience aware of a change of speaker, from the actual narrator to the character of the fox.

There is thus a clear correlation between the semantic-pragmatic structure and the markers formally indicating it. This is especially apparent in the joint use of the SMEs and the connectors: when an SME combines with a connector (see (2a), (7) and (10)), this is a strong indicator of the boundary of a paragraph. By contrast, if one appears without the other — i.e. an SME without connector and vice versa — it does not necessarily indicate a boundary. Note, however, that while the SMEs indicate boundaries between paragraphs, they do not indicate the semantic-pragmatic relationships that exist between these paragraphs. Rather, these relationships must be deduced from the context. This is different with the connectors *nuẓkiṣ* 'thus, so' and *neq^{h}ṣtan* 'then', which express causal and temporal relations.

---

15. Note that by including (1) and (10), we also present clauses beyond the paragraph boundaries of the chosen text segment to show transitions from and to the surrounding paragraphs. It is only by contrasting (1) with (2a) and (9) with (10) that the transitional character becomes apparent.

4.2 Analysis according to SDRT

An analysis of the relationships within the text following a simplified SDRT approach yields nine coordinating and seven subordinating RRs on three hierarchical planes (see Figure 4).

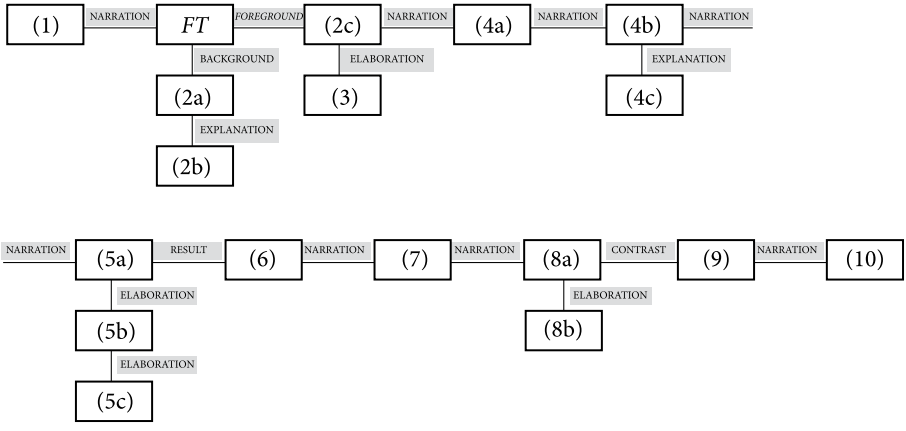


Figure 4. Analysis of the Chipaya text according to an SDRT approach

Of the nine coordinating RRs, NARRATION is the most frequent one with seven instances. This is explicable as NARRATION is probably the one coordinating RR mostly responsible for driving a story forward through a series of "temporally and spatially contiguous" events (Jasinskaja & Karagjosova 2020:7) as the time parameter is intrinsically dynamic. There is one instance of RESULT and CONTRAST each. Subordinating RRs occur seven times, of which ELABORATION is more frequent with four instances, while two relations are classified as EXPLANATION and one represents a BACKGROUND relation. Note the superordinating relations between units (2b) and (2c), (3) and (4a), (4c) and (5a), (5c) and (6) and (8b) and (9).

4.3 Comparison and discussion

Table 2 summarizes and compares the structuring resulting from the SDRT approach and the salience relations.

Table 2 shows that there is no correlation between structuring according to the SDRT approach and the one relying on salience relations. In particular, one might expect the coordinating NARRATION relations to be systematically linked to paratactic transitions as it is there that a new thread of the main storyline begins and NARRATION relations are allegedly closely associated with driving a story for-

Table 2. Comparison of the SDRT and salience relations approach

No	Local – SDRT	Global – Salience relations	
		Transitions	**Formal correlates**
1	Preceding text → (1) Coordination, NARRATION	1 → Summary (paratactic)	Connector
2a	Subordination, BACKGROUND	2 → Shift to king and child (paratactic)	Connector, SME
2b	Subordination, EXPLANATION		
2c	Coordination, NARRATION		
3	Subordination, ELABORATION		
4a	Coordination, NARRATION		
4b	Coordination, NARRATION		
4c	Subordination, EXPLANATION		
5a	Coordination, NARRATION	Parade of the animals	Connector
5b	Subordination, ELABORATION		
5c	Subordination, ELABORATION		
6	Coordination, RESULT		Connector
7	Coordination, NARRATION	3 → Shift to the fox (paratactic)	Connector, SME
8a	Coordination, NARRATION		
8b	Subordination, ELABORATION	hypotactic	SME, SME
9	Coordination, CONTRAST		
10	Coordination, NARRATION	4 → Shift to the armadillo (paratactic)	Connector, SME

ward (see also Section 4.2). However, no such link between paratactic transitions and NARRATION relations can be established: NARRATION relations occur three times at paratactic transitions ((1), (7) and (10)) but five times within a paragraph ((2c), (4a), (4b), (5a) and (8a)).[16] Other RRs do not coincide, either, with the boundaries of the paragraphs and/or their substructuring determined by salience relations. This means that the SDRT approach *per se* does not signal the boundaries of larger discourse units, such as paragraphs.

However, neither the SMEs nor the connectors say anything about the internal semantic-pragmatic cohesion of a paragraph, and, in addition, the SMEs do not indicate the semantic-pragmatic relationships that hold between the units at whose beginning and/or end they occur. For this, other means are required and the SDRT approach appears to be a suitable instrument with which to investigate

16. Note, however, that (5a) marks the internal substructuring of a paragraph.

the semantic relations within paragraphs. Paragraphs and their boundaries, however, are preferably established by means other than SDRT.

The main conclusion to be drawn from this comparison is that we need both for an encompassing analysis of hierarchical discourse structure: an approach like SDRT which captures the relations that hold on the local level but also an approach that makes visible the larger units within a discourse and their respective break-off points, the global level.

5. Conclusions

In this paper, one of our objectives was to come closer to a discourse-functional basis of the analysis of spoken narratives and to enhance our understanding of the semantic-pragmatic side of the coin. We applied a simplified SDRT account (Asher & Lascarides 2003) to determine hierarchical discourse structure and compared the outcome to analyses based on two different text-linguistic approaches (Grimes 1972 & Redeker 2006).

The two selected Samoan and Chipaya texts represent two subgenres of narratives: a personal life story and a traditional story. An interesting result of our analyses which we want to highlight is that they have quite different hierarchical structures. In the Samoan text, the integration of larger units into a whole on a global level reveals a structure which seems contradictory to the assumed sequential order in stories. The presented overlay structure resembles a reversal of narrative sequentiality which, in the visualizations implemented here, would develop in a rightward, horizontal fashion. By contrast, as shown in our analysis, the Samoan story develops by means of a downward-branching, vertical mechanism substantiated mostly by ELABORATIONS. This discrepancy is an important finding, which underlines the need to analyse data from non-Indo-European languages. In addition, some units in the structure have multiple links to preceding units which they elaborate via repetitions. We have captured this structuring effect by the LIST relation. On the whole, we consider the Samoan discourse structure to be an effect of the oral performance. Repetitions are not a gratuitous padding. The speaker is engaged in keeping the hearer on track of what the narrative is about, its point and moral. This speaker activity is relevant to storytelling from an interactive perspective (see Labov & Waletzky 1967; Labov 1972; Bamberg 1994).

In the Chipaya text, on the other hand, we find no correlation between the local SDRT structure and the global one when considering salience relations at discourse transitions. This holds even for the NARRATION relations, which are assumed to develop the main storyline due to the intrinsically dynamic time parameter. However, we would like to point out that despite this apparent mis-

match, the local SDRT structures and the global salience relations complement each other in an important way. Depending on the research question, one or the other approach yields different outcomes, each of which, however, is justified in its own rights. In the introduction of the paper, we underlined the aim of assessing the hierarchical structuring within and beyond paragraphs which is a research desiderate especially with respect to natural spoken data from under-described languages. For determining paragraph structure, we adhered to traditional text-linguistic approaches such as referential movement (see Klein & von Stutterheim 1987). However, for establishing cohesion relations internal to such paragraphs, locally operating approaches like SDRT were required.

We have found that Samoan and Chipaya exhibit additional structures, not attested — or, at least, less frequently so — in Indo-European languages. This study, thus, provides a first glimpse into the cross-cultural variation of the universal communicative practice of narrating. It advances linguistic description based on thorough language documentation in the tradition of Himmelmann (1998) and, what is more, it opens a way to the study of what might be called a linguistic typology of communicative practices. This includes considering formal devices such as discourse markers, referential clitics, deictics and directionals, as well as lexical repetition. For the analysis of such specific discourse-sensitive elements, it turns out to be fruitful to combine analytical approaches. Our particular focus was on paragraphs as units which are problematic for the segmentation and the hierarchical structuring of discourse, and we have shown how these units can be determined specifically on the basis of approaches with a global perspective. Such holistic views on discourse thus seem promising for the analysis of texts rooted in oral performance.

Funding

Part of the research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the Collaborative Research Center 1252 *Prominence in Language*, Project-ID 281511265, which is gratefully acknowledged.Research on Chipaya was also funded by the DFG (grant number: HA 6340/4-1), whose support is gratefully acknowledged.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following glosses are used:

ADD	additive	ASC	ascending
AG	agentive	CLASS.human	human classifier
AL	a-possessive (alienable)	COMPL	completive past tense

CON	connective	LIM	limitative
CONF	confidence	NAL	o-possessive (non-alienable)
CONJ	conjectural	NR	nominalizer
+/ COR	+/ coreferential	NSPEC	non-specific
DC	deictic centre	PERF	perfect
DIR	directional	REASON	reason
EMPH	emphatic	REP	reportative
FT	framing topic;	RESTR	restrictive
GENR	general tense-aspect-mood	+/ SIM	+/ simultaneous
HORT	hortative;	SME	salience-marking enclitic
INC	inclusive	SPEC	specific
IND	indirect	SR	switch-reference
INFORM	informal	V	verbal
LD	locative directional	?	unclear.

References

Adelaar, Willem. 2007. "Threatened Languages in Hispanic South America." In *Language Diversity Endangered*, ed. by Matthias Brenzinger, 9–28. Berlin: De Gruyter Mouton.

Adelaar, Willem, with the collaboration of Pieter Muysken. 2004. *The Languages of the Andes*. Cambridge: Cambridge University Press.

Asher, Nicholas, and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Asher, Nicholas, and Laure Vieu. 2001. "Subordinating and Coordinating Discourse Relations." *Proceedings of the First International Workshop on Semantics, Pragmatics and Rhetoric (SPR'01)*, 57–64.

Asher, Nicholas, and Laure Vieu. 2005. "Subordinating and Coordinating Discourse Relations." *Lingua* 115 (4): 591–610.

Asher, Nicholas, Laurent Prévot, and Laure Vieu. 2007. "Setting the Background in Discourse." *Discours* 1: 1–29.

Bamberg, Michael. 1994. "Actions, Events, Scenes, Plots and the Drama. Language and the Constitution of Part-Whole Relationships." *Language Sciences* 16: 39–79.

Becker, Martin, and Jakob Egetenmeyer. 2018. "A Prominence-based Account of Temporal Discourse Structure." *Lingua* 214: 28–58.

Bublitz, Wolfram. 2006. "It Utterly Boggles the Mind: Knowledge, Common Ground and Coherence." In *Language and Memory. Aspects of Knowledge Representation*, ed. by Hanna Pishwa, 359–386. Berlin/New York: De Gruyter Mouton.

Cerrón-Palomino, Rodolfo. 2006. *El chipaya o la lengua de los hombres del agua*. Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú.

Cerrón-Palomino, Rodolfo. 2009. "Chipaya." In *Lenguas de Bolivia, Vol. I: Ámbito andino*, ed. by Mily Crevels, and Pieter Muysken, 29–77. La Paz, Bolivia: Plural Editores.

Chafe, Wallace L. 1987. "Cognitive Constraints on Information Flow." In *Coherence and Grounding in Discourse*, ed. by Russell Tomlin, 21–52. Amsterdam: Benjamins.

Chiarcos, Christian. 2011. "The Mental Salience Framework: Context-Adequate Generation of Referring Expressions." In *Salience: Multidisciplinary Perspectives on Its Function in Discourse*, ed. by Michael Grabski, Claus Berry, and Christian Chiarcos, 105–139. Berlin: De Gruyter Mouton.

Chiarcos, Christian, Claus Berry, and Michael Grabski. 2011. "Introduction: Salience in Linguistics and Beyond." In *Salience: Multidisciplinary Perspectives on Its Function in Discourse*, ed. by Michael Grabski, Claus Berry, and Christian Chiarcos, 1–28. Berlin: De Gruyter Mouton.

Compes, Isabel. 2000. "Textfunktionen deiktischer Direktionalpartikeln im Samoanischen." In *Räumliche Konzepte und sprachliche Strukturen*, ed. by Christopher Habel, and Christiane von Stutterheim, 119–150. Berlin/Boston: De Gruyter Mouton.

Compes, Isabel. 2002. *Textkonstituierung in gesprochener Sprache: eine Fallstudie zum Verfahren der Verortung am Beispiel samoanischer Narrationen*. München: Lincom Europa.

DobeS. *Dokumentation bedrohter Sprachen, Documentation of Endangered Languages*. https://dobes.mpi.nl/projects/chipaya/ (9 April 2025).

Eberhard, David M., Gary Francis Simons, and Charles D. Fennig. 2020. *Ethnologue: Languages of the Americas and the Pacific*. 23rd ed. Dallas: SIL.

Egetenmeyer, Jakob. 2024. "Textual Development Beyond Temporality and Interactions Between the Adversative Connector *pero* and the *indefinido*." In *Tense, Aspect and Discourse Structure*, ed. by Jakob Egetenmeyer, Sarah Dessì Schmid, and Martin Becker, 297–319. Berlin/Boston: De Gruyter Mouton.

Egetenmeyer, Jakob. In prep. a. Episodes are Determined Backwards. Addressing the Problem of Text-segmentation.

Egetenmeyer, Jakob. In prep. b. The Spanish Additive Connector *además* ('furthermore') in Discourse Structure.

Fleischman, Suzanne. 1986. "'Overlay Structure' in the Song of Roland: A Discourse-Pragmatic Strategy of Oral Narrative." *BLS* 12: 108–123.

Givón, Talmy (ed). 1983. *Topic Continuity in Discourse: A Quantitative Crosslanguage Study*. Amsterdam: Benjamins.

Grimes, Joseph E. 1972. "Outlines and Overlays". *Language* 48 (3): 513–524.

Grimes, Joseph E. 1975. *The Thread of Discourse*. The Hague/Paris: Mouton.

Halliday, Michael A. K. 1967a. "Notes on Transitivity and Theme in English Part I." *Journal of Linguistics* 3 (1): 37–81.

Halliday, Michael A. K. 1967. "Notes on Transitivity and Theme in English Part 2." *Journal of Linguistics* 3 (2): 199–244.

Halliday, Michael A. K. 1968. "Notes on Transitivity and Theme in English Part 3." *Journal of Linguistics* 4 (2): 179–215.

Hannß, Katja. 2008. Uchumataqu. The Lost Language of the Urus of Bolivia. A Grammatical Description of the Language as Documented between 1894 and 1952. PhD Thesis, University of Leiden.

Hannß, Katja. 2021. "Salience and Shift in Salience as Means of Creating Discourse Coherence: The Case of the Chipaya Enclitics." *Pragmatics* 31 (4): 533–559.

Hannß, Katja. 2024. "Uru and Chipaya." In *The Oxford Guide to the Languages of the Central Andes*, ed. by Matthias Urban, 265–310. Oxford: Oxford University Press.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2006. "The Challenges of Segmenting Spoken Language." In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 253–274. Berlin and New York: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2018. "Meeting the Transcription Challenge." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 33–40. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24806

Himmelmann, Nikolaus P., and Beatrice Primus. 2015. "Prominence Beyond Prosody — A First Approximation." In *PS-prominenceS: Prominences in Linguistics. Proceedings of the International Conference*, 38–58. Viterbo: DISUCOM Press.

Himmelmann, Nikolaus P., Meytal Sandler, Jan Strunk, and Volker Unterladstetter. 2018. "On the Universality of Intonational Phrases in Spontaneous Speech — A Cross-linguistic Interrater Study." *Phonology* 35 (2): 207–245.

Hobbs, Jerry R. 1979. "Coherence and Coreference." *Cognitive Science* 3: 67–90.

Jasinskaja, Katja. 2007. Pragmatics and Prosody of Implicit Discourse Relations: The Case of Restatement. PhD Thesis, University of Tübingen. http://hdl.handle.net/10900/46380

Jasinskaja, Katja, and Elena Karagjosova. 2020. "Rhetorical Relations" In *The Wiley Blackwell Companion to Semantics*, ed. by Daniel Gutzmann, Lisa Matthewson, Hotze Rullmann, Thomas Ede Zimmermann, and Cécilie Meier, 1–29. Wiley Online Library.

Johnstone, Barbara (ed). 1994. *Repetition in Discourse. Interdisciplinary Perspectives*. Norwood, NJ: Ablex.

Kehler, Andrew. 2002. *Coherence, Reference, and the Theory of Grammar*. Stanford: CSLI Publications.

Klein, Wolfgang, and Christiane von Stutterheim. 1987. "Quaestio und referentielle Bewegung in Erzählungen." *Linguistische Berichte* 109: 163–183.

Labov, William. 1972. *Language in the Inner City*. Philadelphia: Philadelphia University Press.

Labov, William, and Joshua Waletzky. 1967. "Narrative Analysis: Oral Versions of Personal Experience." In *Essays on the Verbal and Visual Arts*. ed. by June Helm. 12–44. Seattle: University of Washington.

Ladd, D. Robert. 1986. "Intonational Phrasing: The Case for Recursive Prosodic Structure." *Phonology Yearbook* 3: 311–340.

Longacre, Robert E. 1976. *The Anatomy of Speech Notions*. Lisse: The Peter de Ridder Press.

Longacre, Robert E. 1979. "The Paragraph as a Grammatical Unit." In *Syntax and Semantics. Discourse and Syntax*, Volume 12, ed. by Talmy Givón, 115–134. New York: Academic Press.

Longacre, Robert E. 1983. *The Grammar of Discourse*. New York: Plenum Pr.

Longacre, Robert E. 2007. "Sentences as Combinations of Clauses." In *Language Typology and Syntactic Description*. 2nd edition. ed. by Timothy Shopen, 372–420. Cambridge: Cambridge University Press.

Mann, Bill, and Maite Taboada. 2005–2025. *RST Web Site*. Online available at https://www.sfu.ca/rst/ (last access: March 10, 2025).

Mann, William, and Sandra Thompson. 1988. "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization." *Text* 8 (3): 243–281.

McDonnell, Bradley, Andrea L. Berez-Kroeker, and Gary Holton (eds). 2018. *Reflections on Language Documentation 20 Years after Himmelmann 1998*. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24800

Milner, George B. 1993. *Samoan Dictionary*. Reprint of Edition 1966. London: Oxford University Press.

Mosel, Ulrike, and Even Hovdhaugen. 1992. *Samoan Reference Grammar*. Oslo: Norwegian University Press.

Polanyi, Livia. 1988. "A Formal Model of the Structure of Discourse." *Journal of Pragmatics* 12 (5–6): 601–638.

Redeker, Gisela. 2006. "Discourse Markers as Attentional Cues at Discourse Transitions." In *Approaches to Discourse Particles*, ed. by Kerstin Fischer, 339–358. Oxford, Amsterdam: Elsevier.

Renkema, Jan, and Christoph Schubert. 2018. *Introduction to Discourse Studies*. New Edition. Amsterdam: Benjamins.

Sanders, Ted J. M., Vera Demberg, Jet Hoek, Merel C. J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2018. "Unifying Dimensions in Coherence Relations: How Various Annotation Frameworks are Related." *Corpus Linguistics and Linguistic Theory* 17 (1): 1–71.

Sasse, Hans-Jürgen. 1993. "Syntactic Categories and Subcategories." In *Syntax: ein internationales Handbuch zeitgenössischer Forschung / An International Handbook of Contemporary Research*, Vol. 1, ed. by Joachim Jacobs et al., 646–686. Berlin: De Gruyter Mouton.

Serzisko, Fritz. 1992. *Sprechhandlungen und Pausen. Diskursorientierte Sprachbeschreibung am Beispiel des Ik*. Tübingen: Niemeyer.

Serzisko, Fritz. 1995. "The Paragraph in Spoken Narrative Discourse." In *Organization in Discourse. Proceedings from the Turku Conference*, ed. by Brita Wärvik, Sanna-Kaisa Tanskanen, and Risto Hiltunen, 497–508. Turku, Finland: University of Turku.

Tannen, Deborah. 1987. "Repetition in Conversation: Towards a Poetics of Talk." *Language* 63 (3): 574–605.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Librarie C. Klincksieck.

The Language Archive (TLA). "Chipaya Project". https://hdl.handle.net/1839/77f5bf63-956a-4c5a-bb43-c4aeead6c50e (10 April 2025).

Wolf, Florian, and Edward Gibson. 2005. "Representing Discourse Coherence: A Corpus-based Study." *Computational Linguistics* 31 (2): 249–287.

Zeevat, Henk. 2011. "Rhetorical Relations." In *Semantics: An International Handbook of Natural Language and Meaning*, Vol. 1, ed. by Claudia Maienborn, Klaus von Heusinger, and Paul Portner, 946–970. Berlin/New York: De Gruyter Mouton.

PART IV

Diachronic perspectives

CHAPTER 27

Language change in real time

Investigating conversational priming in repetitional responses

Sonja Gipper, Eugen Hill & Pascal Coenen
University of Cologne

Co-present conversation is the primary habitat of human language and thus most probably constitutes an important locus of language change. However, language change is observable only at much larger timescales. How, then, is it possible to study the real-time conversational dimension of language change? In this paper, we argue that language documentation corpora, often covering spontaneous conversational language use, will constitute a crucial data source for such an endeavor, next to historical sources. We report two case studies illustrating the investigation of the mechanism of conversational priming in repetitional responses, where a responding interactant may repeat an innovative form used by the other interlocutor, which may in turn facilitate the spread of this innovation across a population.

Keywords: conversational priming, language change, repetitional responses, spread of linguistic innovations

1. Introduction

One of the critical questions in diachronic linguistics is how an innovative form is transmitted from one person to the next, thus in the long run leading to a spread of the innovation across a population. Given that the primary habitat of language is in face-to-face or co-present (Ameka & Terkourafi 2019) conversation, it can be assumed that the answer to this question may be found there. But language change usually becomes observable only at larger timescales, which makes it difficult to investigate language change at the scale of real-time conversation.

Recent studies have proposed a specific conversational mechanism that could be at play in the propagation of linguistic innovations across a population, the mechanism of conversational priming in repetitional responses (Gipper 2020;

https://doi.org/10.1075/slcs.240.27gip
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

Hill 2022; Dekker et al. 2024). The proposal is based on the observation that interactants may employ repetitional formats in order to perform a range of responsive actions such as answers to polar questions (Holmberg 2016: 62–72; Enfield et al. 2019; Fang 2025). An example of such a repetitional response from German conversational data is given in (1), showing a case where a repeat is employed by speaker D as an answer to B's polar question.[1]

(1) GERMAN [Hoffmann & Himmelmann (2009) – MK-DB]
B: *und* ***das kommt heute?***
and DEM:NOM.N come:3SG.SBJ today
'And that will be broadcast today?'
D: ***das kommt heute***
DEM:NOM.N come:3SG.SBJ today
'That will be broadcast today.'

Now, had interactant B employed an innovative form in her utterance in the question in the first line, this form might have been repeated by D in his response due to priming effects facilitating its use, possibly paired with social motivations discouraging the use of an alternative form. If reiterated various times and in different conversations, such primed uses of innovative forms in repetitional responses may, as we hypothesize, in turn facilitate and accelerate a spread of this innovation across a population. Our proposal is thus consistent with the idea that the spread of linguistic innovations across a population may result from linguistic alignment of interactants during conversation (Trudgill 1986; Auer & Hinskens 2005; Nilsson 2015; Su 2016; Pickering & Garrod 2017).

The crucial question that now arises is what kind of evidence might be adduced to demonstrate the potential influence of conversational priming in language change. Without doubt, the *Documentarist Turn* evoked in this volume inaugurates a new era in the investigation of language change at the conversational scale, with documentation corpora making available records of language use events with different degrees of planning from diverse languages (Himmelmann 1998, 2006). In this paper, we conduct two case studies that demonstrate the use of language documentation corpora in combination with other contemporary and historical data sources in order to investigate our hypothesis.

1. Throughout this contribution, the repeated parts in the repetitional response and the preceding utterance are printed in boldface in the linguistic examples.

2. Conversational priming in repetitional responses

The concept of conversational priming brings together phenomena at different timescales (or causal frames, Enfield 2014): While priming takes place at the level of the psychological processing of language (microgenetic scale), its sequentially structured variant of conversational priming in repetitional responses occurs at the level of social interaction in real time (enchronic scale; see also Enfield 2025). Finally, assuming that conversational priming actually facilitates the spread of an innovation across a population, this is a process working at the diachronic scale of language change (see also Dekker et al. 2024). In the following, we discuss the components of the mechanism in more detail.

The first component is that of a repetitional response, a responsive action that is formatted as a partial, full, expanded, or modified repetition of a preceding utterance by another interlocutor. Repetitional responses can be employed for a variety of actions in conversation (Gipper 2020; Rossi 2020; Brown et al. 2021), among them answers to polar questions (Holmberg 2016: 62–72; Enfield et al. 2019; Fang 2025), requests for reconfirmation (Aldrup 2024; Gipper et al. 2024), and continuer backchannels (Zellers 2021). A partial repeat in an answer to a polar question is exemplified in (2) from Korean, where the predicate is repeated in the affirmative response in the second line. In (3), in contrast, several elements from the question are repeated in the answer.

(2) KOREAN [Kim (2016: 45)]
Hoo: *ani- ssu-nun tey-ka* ***iss-e?***
no write-REL place-NOM exist-INT
'No- there is a place to write {it} in?'
Dan: *e* ***iss-e.***
yes exist-INT
'Yeah there is.'

(3) KOREAN [Kim (2016: 79), simplified]
Nam: *kyay* ***ilpon ka-ss-ta*** ***wa-ss-kwun-a?***
she Japan go-PST-TRANS come-PST-APP-INT
'[I realize] she came back from Japan?'
Gia: *ung* ***ilpon ka-ss-ta*** ***wa-ss-ci.***
yes Japan go-PST-TRANS come-PST-COM
'Yeah {she} came back from Japan.'

Example (4) from Yucatec Maya illustrates the use of a repetitional response in a request for reconfirmation, i.e., an utterance produced as a reaction to news by way of asking the interlocutor to reconfirm what s/he just said (König & Pfeiffer 2024). In the second line of Example (4), speaker L repeats what T has said before

in order for T to reconfirm the previous utterance. T does so by using yet another repeat in the third line (Brown et al. 2021: 261).[2]

(4) YUCATEC [Brown et al. (2021: 262), simplified]
T: ***bin-en in-ch'a'-eh*** *tíiyah*
go-1B 1A-take-SUBJ aunt
'I went to take them, aunt!'
L: *aan* ***bin-ech a-ch'a'-eh***
EXP go-2B 2A-take-SUBJ
'Oh (so), you went to take them.'
T: ***bin-en in-ch'a'-eh***
go-1B 1A-take-SUBJ
'I went to take them!'

The second component of conversational priming is located at the microgenetic scale of language processing: Priming "refers to the phenomenon where one's response to a current stimulus is unconsciously affected by recent exposure to the same or a related stimulus, a mechanism presumed to reflect spreading activation among conceptually — or perceptually — similar structures stored in memory" (Farrell et al. 2012: 206). The first relevant type of priming is that of comprehension-to-production priming, where the employment of a certain form is facilitated through an immediately preceding exposure to this form (Bock 1986; Mahowald et al. 2016; Jacobs et al. 2019). This effect will certainly make it easier for an interlocutor to repeat a linguistic form from the question rather than replacing it with a different form. In addition to this immediate priming effect, replacements may furthermore be heard as corrections or other potentially face-threatening actions (Gipper 2020). One may therefore predict that any given form will be more likely to be repeated rather than replaced by an alternative form in a repetitional response. While priming is mostly investigated in controlled experimental settings, some evidence from experiments involving spontaneous language use is available, suggesting that structures are more likely to be repeated than replaced in answers to content questions (Levelt & Kelter 1982; Chia & Kaschak 2022).

If the interactant now produces a repetitional response and indeed repeats the form in question, a further priming effect arises in addition to the comprehension-to-production priming described above. If this is done, the repetition yields a second instance of the same form. In Example (1) above, for instance, we find two instances of the third person singular verb form *kommt* 'comes'. From the perspective of the interactant who produces the repetitional response (speaker D in Example (1)), the first occurrence of the relevant form pertains to language com-

2. Brown et al. (2021) use the term "request for confirmation" for this practice.

prehension (as it was produced by the other interlocutor, here speaker B), while the second pertains to production (see Bock et al. 2007). This second production use of the form can be assumed to reinforce initial comprehension-to-production priming effects (for a discussion of the current debate on the possibility of self-priming, see Jacobs et al. 2019). Following Dekker et al. (2024: 550) we thus define the mechanism of conversational priming as "the effect of comprehension-to-production priming first facilitating the use of a linguistic form, which then reinforces the priming effect by way of self-priming through production in repetitional responses if the form is repeated rather than deleted or replaced by a different form." In cases where this process is reiterated various times, cumulative priming effects (Kaschak et al. 2014) may lead to the interactant's familiarization with the form resulting in its integration in the interlocutor's linguistic repertoire. If many members of a community become familiarized with an innovation in this way, this may facilitate the spread of the innovation across a population. Most likely, the condition for such a spread is that the innovative form has some advantage over the old form, for instance in terms of ease of pronunciation or of social prestige (Blythe & Croft 2012; Dekker et al. 2024).

3. In search of empirical evidence

While priming has been suspected to be a central player in language change (Jäger & Rosenbach 2008; Pickering & Garrod 2017; Baumann & Sommerer 2018; Kootstra & Şahin 2018), it is notoriously difficult to empirically investigate this. As Mair (2017: 191) points out: "Priming studies work with time frames of seconds and minutes, whereas historical linguistics deals with centuries and millennia". Due to this difficulty of observing the potential long-term effects of conversational priming in real time, Dekker et al. (2024) conducted a simulation in the form of an agent-based model in order to investigate the possibility of an effect of conversational priming in language change. They find that conversational priming speeds up the processes of convergence and spread of innovation across a population. This suggests that an effect of conversational priming in real life is in principle possible. Although computer simulations may be very helpful as a first plausibility test, no linguistic hypothesis can be seen as sufficiently supported until its effects are demonstrated on the basis of real-world linguistic data.

How, then, is it possible to investigate the hypothesis that conversational priming through repetitional responses plays a role in language change? In the following sections, we discuss possible ways forward, focusing on data sources (3.1) and the possibility of investigating the hypothesis indirectly on the basis of certain grammatical asymmetries engendered by repetitional responses (3.2 and 3.3).

3.1 Data sources

Conversational priming in repetitional responses is a phenomenon that is observable at the real-time enchronic scale of co-present conversation. Therefore, the first place to look for evidence of the role of conversational priming in language change is recordings of everyday conversation in many different languages. For investigating our hypothesis, we would ideally be able to identify a language change in progress in the language under investigation. If this change affects a variable that is relatively frequent in language use, we may track its trajectory of change in real time, including both its overall behavior as well as its behavior in repetitional responses. If no change can be detected, any kind of variation may serve as a proxy. For this purpose, the ideal data source would be a longitudinal corpus of spontaneous conversations covering an ongoing linguistic change from the very first employments of the new variant to the completion of the change, i.e., the point in time when the new variant has spread across the population. Fortunately, the recent *Documentarist Turn* has put us in a position to pursue such investigations. Language documentation projects around the world have yielded numerous language documentation corpora of a broad variety of languages. As envisioned by Nikolaus Himmelmann in his seminal 1998 paper, a language documentation corpus should ideally contain a broad range of different types of communicative events, ranging from completely unplanned exclamations such as *ouch!* to discourse genres that involve a high degree of planning, such as ritual discourse. More or less in the middle, we find conversational genres, including chat, discussion, and interview (Himmelmann 1998: 180). These conversational data will lie at the heart of future empirical investigations of language change at the enchronic scale, alongside historical sources and corpora capturing naturalistic language use in more well-described languages (see also Bauer et al., this volume).[3]

But language documentation corpora have more to offer. Clearly, spontaneous conversations are not the only data type that may be relevant for investigating our hypothesis. Traditional narratives and folk stories, for instance, with their tendency to preserve more conservative historical tiers of a language (see Reinöhl, this volume), will constitute an important data source for detecting ongoing or completed changes in a language. Moreover, they may preserve historical interactional practices, which may surface, for instance, in represented conversations in narratives or in ritual interactions. This means that any given language documen-

3. A case in point is the FOLK corpus (e.g. Kaiser 2018; Reineke et al. 2023), a stratified corpus of German interactional data in a broad range of situations, which will soon cover a time span of 20 years.

tation corpus is likely to contain data that are relevant to investigating repetitional responses and their potential role in language change. In this way, language documentation corpora live up to the original purpose of language documentation to build "a lasting, multipurpose record of a language" (Himmelmann 2006: 1).

3.2 The person marking asymmetry

In this section, we flesh out a possibility of finding indirect evidence for the influence of repeats on the spread of innovations. Our approach hinges on the observation that repeats engender formal asymmetries, where certain grammatical forms are typically repeated exactly in a response whereas others are usually not. In languages with distinct person marking, for instance, a question in the third person will be answered with a repeat in the third person as well, leading to an exact repetition. In an answer to a question in the second person, however, a deictic shift to the first person will be necessary in the answer, leading to the use of a distinct verb form (Hill 2022). Compare the following question-answer sequences from Yurakaré (isolate, Bolivia). In (5), the predicate of the question is in the second person singular, whereas in (6), the predicate is in the third person plural.

(5) YURAKARÉ [van Gijn et al. (2011) — 250906_convIII]
A: ***amala-m*** *latiji komadre=la*
come-2SG.SBJ then comadre=COM
'Then you came, comadre?'
P: ***amala-y*** *komadre*
come-1SG.SBJ comadre
'Then I came, comadre.'

(6) YURAKARÉ [van Gijn et al. (2011) — Conversation-NL]
E: ***dula-shta-w*** *reunion?*
do-FUT-3PL.SBJ meeting
'Will they have a meeting?'
S: ***dula-shta-w=ya***
do-FUT-3PL.SBJ=REP
'They will have a meeting (it is said).'

When the question is asked in the second person, the answer will usually have to be given in the first person. As a result, there is no exact repetition of the predicate in (5). In contrast to this, the predicate is exactly repeated when the question is asked in the third person, as in (6). If conversational priming is indeed a vehicle for the spread of new grammatical forms from one interactant to another, we predict that these synchronic asymmetries are also reflected in the diachronic development of these forms. More specifically, we assume that innovations in the

third person, which has to be repeated exactly, may spread faster than those in the first and second person singular, which are not usually repeated exactly, all other things being equal.

If conversational priming indeed facilitates the spread of novel grammatical forms, this should be observable in the history of Indo-European languages, for repeats as default answers can be found in the oldest attested stages of several languages (Hackstein 2024), for instance in Vedic Sanskrit (Lühr 2018: 174–175) or in Latin (Brown et al. 2009: 515–516; Hackstein 2021; Potočnik 2023). As expected, the relevant asymmetries between persons are observable. This can be illustrated by the examples from Vedic Sanskrit in (7) and (8).[4]

(7) VEDIC SANSKRIT [Śatapatha-Brāhmaṇa 10.3.4.2]
véttha catvā́ri mahatā́m mahā́ntī3 íti
véttha *catvā́ri mahatā́m mahā́ntī3 íti*
know:PRF.2SG.SBJ four:ACC.PL.N great:GEN.PL.N great:ACC.PL.N QUOT
véda bho3 íti hovāca
véda *bho3 íti ha uvāca*
know:PRF.1SG.SBJ INTERJ QUOT PART say:PRF.3SG.SBJ
'"Knowest thou the four great ones of the great?" "I know them, sir," he said.'

(8) VEDIC SANSKRIT [Śatapatha-Brāhmaṇa 11.6.1.4]
sá hovācā́stī́ha
sá ha uvāca ***ásti*** *íha*
DEM:NOM.SG.M PART say:PRF.3SG be:PRS.3SG.SBJ here
prā́yaścittī3rítyástī́ti
prā́yaścittī3ḥ íti ***ásti*** *íti*
atonement:NOM.SG.F QUOT be:PRS.3SG.SBJ QUOT
'He said, "Is there no atonement for this?" "Yes, there is," they replied.'

We propose that it may be possible to observe the impact of repeats on the spread of innovations by looking at the components of the grammar of a given language where such asymmetries arise. Concretely, we expect forms that are regularly repeated exactly in repeats to change faster, on average, than those forms that are not regularly repeated. In our case study in Section 5, we use historical data from Lithuanian to show that our expectations indeed bear out for the preterite tense of Lithuanian *léisti* 'to let, release, let go'. The historical data suggest that it was indeed the third person form that changed fastest in the 16th–17th centuries, whereas the first person singular lagged behind.

4. The Vedic text of (7) and (8) is adopted from Gippert et al. (2016). In the second line, we have resolved the sandhi forms. The English translation is adopted from Eggeling (1882–1900: IV, 333; V, 109).

The person asymmetry described above is, moreover, particularly well-suited to test a potential contribution of conversational priming in repetitional responses to language change, as its predictions are opposed to those of the frequency distribution of the different persons (see also Dekker et al. 2024). It is well known that in inflectional change, more frequent words and word forms are less prone to change and/or resist the change for a longer time than the less frequent ones (see, most recently, Janda et al. 2010; De Smet & Van de Velde 2019; Krause-Lerche 2022). Given that previous studies suggest that the third person singular is the most frequent in spoken language at least in some languages (e.g., Scheibmann 2001 for American English; Seržant & Moroz 2022 for Russian), from this follows that frequency and conversational priming make opposite predictions: Whereas frequency predicts that the most frequent form, i.e., the third person singular, should be the most conservative, conversational priming predicts the exact opposite, that the third person should be the most innovative (see also Dekker et al. 2024).

This person marking asymmetry is, of course, only relevant in languages with person marking systems that formally distinguish the first, second, and third person. An absence of the contrast between first- and second-person singular can be observed, for instance, in languages that possess egophoric marking. Egophoricity is a linguistic category that is susceptible to differences in knowledge and involvement, whose distribution is sensitive to person, role in the speech act, and sentence type. Thus, a common distribution of an egophoric marker is with a first-person subject in a declarative sentence and with a second person subject in an interrogative sentence (San Roque et al. 2018: 2). This means that in a polar question–answer sequence concerning the addressee, the same egophoric EGO marker will be used in the question and the answer, as in Example (9) from Cha'palaa (Barbacoan, Ecuador) where the EGO marker is used for agentive involvement and the NON-EGO marker for other kinds of involvement in a situation (Floyd 2018: 269). For languages with egophoricity, thus, conversational priming will generate different predictions which will be an intriguing point of comparison in future research.

(9) CHA'PALAA [Floyd (2018: 296)]
A: ***panda ʃi-yu***
food eat-EGO
'(Did you) eat?'
B: ***panda ʃi-yu***
food eat-EGO
'(I) ate.'

To sum up, we propose that the person marking asymmetry allows for an indirect investigation of potential effects of conversational priming in repetitional re-

sponses. In particular, for languages with distinct person marking, the predictions of conversational priming oppose those of frequency, which makes this asymmetry an ideal testing ground for our hypothesis. Crucially, this kind of evidence is indirect, in that it does not necessarily consist of conversational data. The idea is that once an innovation has reached a certain spread facilitated through conversational priming in repetitional responses, this will become observable outside responses as well. The advantage of this type of evidence is that it can be found in any kind of corpus.

3.3 Further asymmetries

In some languages, there are language-specific conventionalized formats of repeats that do not contain the verbal predicate. A case in point is Slovenian, where a standard means of answering a polar question is in the form of a clitic pronoun (Dvořák 2007).[5] If the question already contains such a clitic, the answer can be analyzed as a repeat (see also Holmberg 2016: 62) as illustrated in (10).

(10) SLOVENIAN [Dvořák (2007: 210)]
A: *A **mu** verjameš?*
Q CL.3.M.DAT believe:2SG.SBJ
'Do you believe him?'
B: ***Mu**.*
CL.3.M.DAT
'I do.'

Object clitics are not the only pronouns that can be repeated in order to answer a polar question. In several languages, answers to polar questions contain a response particle combined with a subject pronoun. In cases where the polar question contains a corresponding subject pronoun, the pronoun in the answer constitutes a repeat. This pattern is illustrated in (11)–(13) by data from Middle High German, Middle Low German and Middle English.

(11) MIDDLE HIGH GERMAN [Steinhoff (1995: 744)]
*»Ist **er** hinweg«, sprach der konig.*
be:PRS.3SG.SBJ 3SG.M.NOM gone say:PRT.3SG.SBJ ART:NOM.SG.M king:NOM
*»Ja **er**«, sprach myn herre Gawan, […]*
yes 3SG.M.NOM say:PRT.3SG.SBJ 1SG.GEN HON PN
'"Is he gone?" said the king. "Yes" said my Lord Gawan, […]'

5. This applies to polar questions containing a verb in the present tense and an object in the dative or accusative. Dvořák (2007) gives further specifications for the context for these answers, which we do not discuss in detail here.

(12) MIDDLE LOW GERMAN [ReN-Team (2021: Kortw. Hist. Laurin 65r,29)]
He sprack / hebbe ***gy*** *na my*
3SG.M.NOM say:PRT.3SG.SBJ have:2SG.SBJ.POL 2SG.POL.NOM after 1SG.DAT
gesandt ? Ja ***ick*** */ so sprack de koning*
send:PP yes 1SG.NOM so say:PRT.3SG.SBJ ART:NOM.SG.M king:NOM.SG.M
thohandt
right_away
'He said: "Have you sent for me?" "Yes" said the king right away.'

(13) MIDDLE ENGLISH [Morris (1876: 782)]
"[…] Wil ***yee*** *discipls his bicum?"* ***"We,"***
will:2PL.SBJ 2PL.NOM disciple:PL 3SG.M.POSS become we
coth þai, […]
say:PRT.3PL.SBJ 3PL.NOM
'"[…] Will you become his disciples?" "Yes" they said, […].'

Note that the above described strategy is attested for both asymmetric and symmetric pronoun use. In the former, the question contains a second person subject and the answer, correspondingly, a first person subject — see Examples (12) and (13). Example (11) illustrates a symmetric pronoun use, where both the question and the answer contain the same third person pronoun, *er* 'he'. This conventionalized use of pronouns in answers to polar questions thus bears the same asymmetry as described in Section 3.2 for verbal person marking. Similar strategies are observable as early as in ancient Indo-European languages (cf. Hackstein 2021; 2024), but in contemporary German and English conversational data, such repeats of subject pronouns are no longer commonly found.

However, not only the first/second persons are subject to asymmetries. In the third person, asymmetries may arise, for instance, with demonstrative pronouns. In Yurakaré, the demonstrative *ana* is mostly used for the function of indicating a referent to the addressee, while *ati* virtually exclusively indexes referents that have been introduced into the discourse previously (Gipper 2017). Hence, in Yurakaré, in the case of co-reference the sequence *ana — ana* is strongly dispreferred in sequences containing repetitional responses, whereas *ana — ati* is the unmarked pattern, as exemplified in (14).

(14) YURAKARÉ [Gipper & Ballivián Torrico (2014) — SocCog-YUZ108-2]
A: ***ana=ja ultimo-shta-ø lëshie-w=ra***
DEM=REA last-FUT-3SG.SBJ two-3PL.SBJ=right
'This is the last two [pictures].'
B: ***ati=ja ultimo-shta-ø lëshie(-w)***
DEM=REA last-FUT-3SG.SBJ two-3PL.SBJ
'That is the last two.'

The abovementioned asymmetries all constitute cases of indexical phenomena, where the asymmetry is caused by an indexical shift in the asymmetric cases. For these cases, our predictions are parallel to that derived from the person-marking asymmetry described in Section 3.2: We expect a change to spread more easily and thus faster in cases where the respective form is repeated in the response than in cases where the form is replaced due to indexical shifts.

In addition to the indexical cases described above, there are other conventionalized repeat formats, one of which is found in dialects of Lithuanian. This strategy is confined to questions with verbs containing so-called preverbs. Lithuanian preverbs are proclitics on verbs that modify the semantics of the verb to which they are attached. In dialects of Lithuanian, a polar question containing a verb with a preverb may be answered in the affirmative by repeating the preverb alone, at least for verbs in the past and future tenses. The *Lietuvių kalbos žodynas* (LKŽ; Kruopas et al. 1968–2002), which is the major source of information on 20th c. spoken dialects, attests preverb-repeats predominantly for the northern part of Lowland Lithuanian, but also for some northern Highland Lithuanian local dialects. Zinkevičius (1966: 335) confirms this and states that preverb-repeats may also occur in the south of both dialect areas. For further evidence, see also Senn (1966: 305–306) and Aleksandravičius (2011: 27, 122, 127, 261, 266, 315, 319, 396, 467). Historically, the available evidence (Ostermeyer 1791: 163; Lepner 1690/1744: 113) attests the use of preverb-repeats as early as in the late 17th c. in western Highland Lithuanian and, at least since the mid-19th c., also in northern Lowland Lithuanian. (15) and (16) offer a selection of representative examples.

(15) LITHUANIAN [LKŽ I:350]
Ar ***at****=vedė* *arklį?* *–* ***At.***
Q toward=lead:PRT.3.SBJ horse:ACC.SG toward
'Did he bring a horse?' — 'Yes.'

(16) LITHUANIAN [LKŽ IV:5]
Ar ***su****=vežei* *rugius?* *–* ***Su.***
Q with=drive:PRT.2SG.SBJ rye:ACC.PL with
'Did you collect the rye?' — 'Yes.'

We assume that these preverb-repeats existed alongside repeats of verbs in cases where the question did not contain a preverb. For the Lithuanian preverb-repeats, our hypothesis then predicts that any linguistic innovation should be able to spread faster in verbs without preverbs, as these are repeated in repetitional responses, while for verbs with preverbs, only the preverb but not the verb gets repeated.

The above-described phenomena capture only a fraction of the potential cross-linguistic diversity of conventionalized repeat formats. Only a profound

knowledge of the forms and functions of repeats across languages will allow us to identify potential linguistic variables that may be subject to a facilitation of change via repeats. Thus, mapping out the cross-linguistic variability space of repetitional response is a crucial endeavor in future research. Again, language documentation corpora — which were conceived from the outset as "a primary data pool for typological inquiry" (Himmelmann 2016: 474) — provide an ideal data source for investigating the forms, functions, and frequencies of repeats in conversation in a broad range of languages and thus for arriving at a typology of repetitional responses in conversation.

4. Case study I: The occurrence rate of repeats in conversation

Repeats in different interactional functions are attested in a diverse range of modern and historical languages, among them, in addition to the languages discussed in this paper, Welsh (Jones 1999: 43–128), European and Brazilian Portuguese (Thomas 1969: 242–243; Martins 2013; Rosemeyer & Schwenter 2019; Harjunpää & Osterman 2023), Estonian (Keevallik 2010), Finnish (Sorjonen 1996), Low German (Weber 2024), Central Alaskan Yupik (Miyaoka 2012: 169), and the Khoisan language ǂĀkhoe Haiǁom (Hoymann 2010). For testing our hypothesis that repetitional responses facilitate the spread of linguistic innovations, it is crucial to know the occurrence rate of repeats in everyday conversation, as the hypothesis becomes less plausible if repeats are very infrequent. Previous research suggests that some language communities make more pervasive use of repetitional responses than others (Enfield et al. 2019; Gipper et al. 2024).[6] Moreover, in some languages (possibly those with less pervasive use of repeats), repeats tend to convey more marked or even competitive actions, whereas in others repeats regularly convey more unmarked and affiliative actions (Brown et al. 2021; Gipper et al. 2024; Fang 2025). However, studies usually focus on one particular action such as affirmative answers to polar questions (Enfield et al. 2019), or do not pursue a quantitative approach (Brown et al. 2021). Consequently, we lack comparative quantitative studies of the overall occurrence rate of repeats in conversation. In this section, we report results from a pilot study attempting a comparison of the conversational frequencies of repeats in two languages, German and Yurakaré, hypothesizing on the basis of previous research (Deppermann et al. 2024; Gipper 2024; Gipper et al. 2024) that repeats will be less frequent in German than in Yurakaré.

6. See also Stivers (2010: 2779), Rossano (2010: 2770), Enfield (2010: 2662), Hayashi (2010: 2700) and Hoymann (2010: 2734) on the ratio of repeats in polar answers in American English, Italian, Lao, Japanese and ǂĀkhoe Haiǁom, respectively.

The Yurakaré people constitute one of the Indigenous groups of the Plurinational State of Bolivia. The Yurakaré language is spoken by roughly 1,600 people (INE 2015: 32). Most speakers of the language are bilingual with Spanish, and children nowadays mostly grow up as Spanish monolinguals (Plaza Martínez et al. 2011: 243–245). Due to this break in intergenerational transmission, the Yurakaré language is highly endangered. The first documentation project on the language was carried out from 2006 to 2011 by the linguist Rik van Gijn, the anthropologist Vincent Hirtzel, the Yurakaré speaker and language expert Jeremías Ballivián Torrico, and the first author of this contribution. This project entitled *The Documentation of Yurakaré* was funded by the DoBeS Initiative of the Volkswagen Foundation. The data collected within the project are archived at *The Yurakaré Archive* (van Gijn et al. 2011). For further information on the project, the reader is referred to Hirtzel (2021).

For the pilot case study conducted in this section, three conversational dyads from *The Yurakaré Archive* were selected, featuring one male and four female speakers (one of the speakers participates in two of the recordings). For German, we used three conversational dyads from the MünsterKorpus, a corpus containing casual conversations among fellow students and friends dating from 2009 (Hoffmann & Himmelmann 2009). The German dyads feature three female and three male speakers. The data for both languages are roughly comparable regarding the activity type, as all dyads constitute mundane conversations without any other main activities (such as meal preparation or other work) ongoing in parallel.[7] All recordings constitute what Himmelmann (1998) calls "staged communicative events", i.e., conversations that were carried out for the purpose of the recording. Regarding the speakers' age, however, the data are not fully comparable. In the German sample, the speakers are roughly in their twenties, whereas in the Yurakaré data set, one of the speakers is younger than 20 (according to the available information), two speakers are in their forties, and two in their sixties. Variation in the use of repeats according to age will be an interesting avenue for future research. Despite this, we believe that the case study can give us a first impression of the occurrence rate of repeats in the two languages under investigation.

For this case study, all repetitional responses were tagged in selected stretches of talk in the respective dyads, regardless of their interactional function. We in-

7. In two of the Yurakaré dyads, the interactants at some point start engaging in the telling of mythological narratives. In total, roughly six minutes represent this activity type out of the total of 45 minutes sampled. In both cases, this activity type was freely chosen by the interlocutors. While it may not be fully comparable to other activities such as first-person narratives or back-and-forth discussions, it is still similar in that it is fully interactive, involving active participation by the recipient interlocutor.

cluded full repeats, partial repeats, and repeats with some modification or expansion. For the German data, 30 minutes per dyad were coded, while for the Yurakaré data we coded 15 minutes per dyad. All data were coded in ELAN (Wittenburg et al. 2006).

The occurrence rates of repeats in the German and Yurakaré conversational dyads are summarized in Table 1. In all Yurakaré dyads and for all Yurakaré speakers, repeats show a higher average occurrence rate per minute than in the German dyads. While for the Yurakaré speakers, the rate of repeat production ranges from one to 3.5 repeats per minute, the six German speakers produce a repeat with a rate of 0.3 to 0.5 per minute. The data in Table 1 thus support the hypothesis that repeats are more frequent in Yurakaré than in German conversation.

Table 1. Absolute and normalized (per minute) frequencies of repetitional responses (RRs) in German and Yurakaré conversational dyads

German			Yurakaré		
Speaker	**RRs**	**RRs per min.**	**Speaker**	**RRs**	**RRs per min.**
DL	9	0.3	A	52	3.5
B	16	0.5	M	48	3.2
U	12	0.4	A	46	3.1
V	9	0.3	P	42	2.8
L	13	0.4	N	16	1.1
DD	14	0.45	Y	15	1
	Total: 73	**Average: 0.4**		**Total: 219**	**Average: 2.45**

With respect to our hypothesis, the findings in Table 1 raise questions regarding the possible relationship between the occurrence rate of repetitional responses and their potential impact on language change. One possible hypothesis is that languages where repeats are used with higher frequencies have the capacity of changing faster than languages with lower rates of repeats, at least in those grammatical domains that get repeated exactly in line with the asymmetries introduced in Sections 3.1 and 3.2. One may also question the relevance of our hypothesis for those languages with low rates of repeats: Given that speakers of German produce a repeat only 0.4 times per minute, can repeats still be relevant for language change? Regarding this question, we have to keep in mind that humans are engaged in conversation for various hours a day. Recent large-scale research conducted in Spain suggests that on average, people engage in face-to-face, co-present interaction roughly 220 minutes a day, with an additional 47 minutes of phone conversation (Marijuán et al. 2019: 82–83). To the extent that speakers of German

behave in similar ways, and taking an average of 0.4 repeats produced per speaker per minute in German conversation (Table 1), this results in roughly 107 repeats produced by an average speaker of German each day. While this is only a very rough and preliminary calculation, the results from this pilot case study suggest that even in languages with comparatively low rates of repetitional responses such as German, interactants produce a considerable amount of repeats a day. We propose that this is enough for repeats to have an impact on the spread of innovations, in particular for frequent forms such as verbal person marking.

5. Case study II: A change in the preterite tense of Lithuanian *léisti* 'to let, release, let go'

In our second case study, we conduct an indirect investigation of the expected effects of conversational priming based on the asymmetries described in Sections 3.2 and 3.3. We focus on a change in the preterite of Lithuanian that took place in the 16th and 17th centuries. Alongside its neighbor Latvian and the now extinct language Old Prussian, Lithuanian belongs to the Baltic sub-branch within the Balto-Slavonic branch of Indo-European. At present, Lithuanian is spoken nearly exclusively within the borders of the Republic of Lithuania. It is, however, known that during the centuries between the first written records in the mid-16th c. and the re-establishing of the Republic of Lithuania in the late 20th c. Lithuanian covered a larger area. Apart from Lithuania itself, this area also embraced adjacent territories of present-day Belorussia and Poland, as well as, most importantly, large parts of what used to be the Duchy of Prussia (from the late 18th c. on the Province of East Prussia belonging to the Kingdom of Prussia and, subsequently, Germany).

In present-day Lithuanian, polar questions are typically answered in the affirmative by using the particle *taip* 'so; yes' or by repeating a part of the question, often its predicate. A combination of both, using the particle and repeating the predicate, is also common. The use of a repeat is illustrated in (17).

(17) LITHUANIAN [Corpus of Contemporary Lithuanian Language (Utka et al. 2017), subcorpus "colloquial speech", SK_2004]

*Ar **girdėjai** mane Alfredai?*
Q hear:PRT.2SG 1SG.ACC PN:VOC.SG
'Did you hear me, Alfredas?'
– ***Girdėjau**.*
hear:PRT.1SG
'Yes.'

Given the use of repetitional responses in spoken Lithuanian, we expect that the evolution of its inflection is subject to the person asymmetry introduced in Section 3.2. Accordingly, we predict that a given inflectional change would first affect the third person, before the inflectional forms of first and second person follow. But there is also a second prediction that we can test based on the preverb-repeat answers described in Section 3.3. Answers to polar questions containing a preverb repeat only the preverb and not the verb, while in answers to polar questions not containing a preverb, the verb is repeated. We may thus expect that verb forms *not* containing a preverb change faster, as they are repeated in repetitional responses. In this section, we test whether these theoretical predictions are borne out by the data on Lithuanian dialects collected between the start of written records in the mid-16th c. and the present days.

As a specific case, we look at the preterite tense of Lithuanian *léisti* 'to let, release, let go'. We do so for the following reasons. This verb is frequent and well attested in Lithuanian since the very beginning of its written records. According to LKŽ (VII:244–295), *léisti* is broadly used with a variety of preverbs; at least its use with *ap=* 'around', *at=* 'toward', *į=* 'in', *iš=* 'out', *nu=* 'from', *pa=* 'along', *per=* 'over', *pra=* 'in front', *pri=* 'by', *su=* 'with', and *už=* 'behind' is well attested for most dialects. Moreover, Lithuanian has only two different patterns of preterite tense formation: the *ė*-preterite and the *o*-preterite. These can be most clearly distinguished in the third person, which in Lithuanian has the same form in all three numbers. The PRT.3 of an *ė*-preterite ends in *-ė*, that of an *o*-preterite ends in *-o*. An overwhelming majority of Lithuanian verbs form either an *ė*- or an *o*-preterite in all contexts. Now, Lithuanian *léisti* belongs to a small group of verbs which are attested with both an *ė*-preterite, i.e., PRT.3 *léidė*, and an *o*-preterite, i.e., PRT.3 *léido* (Zinkevičius 1966:355; Ambrazas et al. 1991:100–102). Both paradigms are given in Table 2 (note that the reflexive particle in verb forms containing a preverb, such as *at=* 'toward', may appear between the preverb and the verb root). In the 20th c., the *ė*-preterite of *léisti* was in use in local Highland Lithuanian sub-dialects scattered around Šiauliai in the west and Panevėžys, Kupiškis, Zarasai, and Molėtai in the eastern part of Lithuania. Besides that, PRT.3 *léidė* is attested in the archaic Highland Lithunian dialect island of Zietela in Belorussia. All other Lithuanian dialects, including the present-day standard language, exhibit the *o*-preterite with PRT.3 *léido*. This picture suggests that Lithuanian *léisti* is in a process of transition from an inherited *ė*-preterite (PRT.3 *léidė*, etc.) to a more recent *o*-preterite (PRT.3 *léido*, etc.). The inherited character of the former is supported by closely related Latvian. In those dialects of Latvian which preserve the distinction between both types of preterite tense formation, the verb *laîst* 'to let' exhibits the Latvian counterpart of the Lithuanian *ė*-preterite (Endzelin 1910:29; Reķēna 1998:551–552; Strautiņa 2007:67, 147).

Table 2. Preterite paradigms for *léisti*

	ė-preterite			*o*-preterite		
		Reflexive			Reflexive	
PRT.1SG	*léidžiau*	*léidžiau=s(i)*	*at=si=léidžiau*	*léidau*	*léidau=s(i)*	*at=si=léidau*
PRT.2SG	*léidei*	*léidei=s(i)*	*at=si=léidei*	*léidai*	*léidai=s(i)*	*at=si=léidai*
PRT.1DU	*léidėva*	*léidėvo=s*	*at=si=léidėva*	*léidova*	*léidovo=s*	*at=si=léidova*
PRT.2DU	*léidėta*	*léidėto=s*	*at=si=léidėta*	*léidota*	*léidoto=s*	*at=si=léidota*
PRT.1PL	*léidėm(e)*	*léidėmė=s*	*at=si=léidėm(e)*	*léidom(e)*	*léidomė=s*	*at=si=léidom(e)*
PRT.2PL	*léidėt(e)*	*léidėtė=s*	*at=si=léidėt(e)*	*léidot(e)*	*léidotė=s*	*at=si=léidot(e)*
PRT.3	*léidė*	*léidė=s(i)*	*at=si=léidė*	*léido*	*léido=s(i)*	*at=si=léido*

When exactly the transition from PRT.3 *léidė* to PRT.3 *léido* started in Lithuanian cannot be determined with certainty. The PRT.3 *léido* is attested twice, spelt ⟨leida⟩, as early as in the second part of the *Gesmes Chriksczoniskas* by Martynas Mažvydas (1510–1563), which was printed in the year 1570 (Gerullis 1923: 436, 469).[8] Since the composition of the book must have preceded its author's death, the earliest attestation of PRT.3 *léido* thus postdates the earliest known traces of written Lithuanian, glosses in a Latin book (see Subačius et al. 2010), by no more than some 30 to 40 years. The innovative *o*-preterite of *léisti* is frequently attested in the writings of Johannes Bretke (1536–1602). In his *Postilla*, a collection of Lithuanian sermons published in 1591, the *o*-preterite of *léisti* occurs no less than 53 times, while the *ė*-preterite of *léisti* does not occur at all.[9] It follows that in parts of the Lithuanian language area the innovative *o*-preterite of *léisti* evolved no later than in the mid-16th c. and, by the end of this century, it completely ousted the inherited *ė*-preterite in the speech of at least one speaker (i.e. Bretke).

However, in several sources which are only slightly younger both the *ė*-preterite and the *o*-preterite of *léisti* are found side by side. The earliest of such sources is the *Kathechismas* translated from a Polish original and published in 1595 by Mikalojus Daukša (appr. 1527–1613), who was born in or near Kėdainiai (western Highland Lithuanian) and worked as a catholic pastor in Varniai (southern Lowland Lithuanian). The language of Daukša's writings is western Highland Lithuanian of the Kė-

8. In Mažvydas writings, unstressed word-final *-o* is spelt ⟨o⟩ or ⟨a⟩ (Zinkevičius 1978a: 39; 1978b: 139–140) while *-ė* appears as ⟨e⟩ or ⟨ie⟩ (Zinkevičius 1977: 360). It follows that ⟨leida⟩ can only represent *léido*.

9. In this text, unstressed word-final *-o* is nearly always spelt ⟨a⟩ (cf. the very frequent ⟨buwa⟩ for *bùvo* 'was, were'), while *-ė* is always ⟨e⟩ (as in ⟨numire⟩ for *nùmirė* 'died'). It follows that ⟨leida⟩ must represent *léido*.

dainiai region. In Daukša's *Kathechismas*, the PRT.3.REFL *léido=s* (in *su=si=léido=s*, printed ⟨susiłaidoś⟩ 130_9), which is a clear *o*-preterite form, is accompanied by the *ė*-preterite form PRT.1SG.REFL *léidžiau=s* (in ⟨priłeidźeuś⟩ 187_{18}).[10] Unfortunately, all other preterite tense forms of *léisti* belong to the ambiguous PRT.2SG.

More data are found in the *Postilla catholicka*, a collection of sermons translated from Polish by the same author and published four years later, in 1599. In this text, which is much longer than the *Kathechismas*, 139 finite preterite tense forms of *léisti* are attested. Their distribution over the paradigm is shown in Table 3, omitting the ambiguous PRT.2SG.[11] The data show that in this text, the *o*-preterite of *léisti* is much more frequent than the *ė*-preterite. Remarkably, all 5 attestations of the PRT.1SG belong to the inherited *ė*-preterite.[12]

Taken together, Daukša's writings indicate that he, while translating the Polish original and/or correcting the translation, considered the *o*-preterite more appropriate in the PRT.3 and the *ė*-preterite in the PRT.1SG. This is exactly the picture predicted by the person asymmetry described in Section 3.2, according to which the 1SG, a form typically not repeated exactly in answers, should be affected by change later than its third person counterpart.

As for the asymmetry between preterite tense forms with and without preverbs, the results are not clear. In Daukša's writings, the verb *léisti* hardly ever occurs without a preverb. Out of all 10 simplex forms, only 1 pertains to the inherited *ė*-preterite (10%). Regarding the preverb forms, 16 out of a total of 120 are in

---

10. Here and in the following the number refers to the page number in the source and the index to the line on this page.

11. The data have been taken from the index by Kudzinowski (1977). Each attestation has been checked with the reproduction of the original text in Palionis (2000); the mistakes in Kudzinowski's index (such as the wrong assignment of PRT.3.REFL *léido=s*(*i*) translating Pol. *udał*(*a*) *się* to the PRS tense in I:410) have been corrected.

12. In theory, the *ė*-preterite forms traditionally attributed to *léisti* in Daukša's writings might also belong to a different verb, namely the iterative of *léisti*, which is *láidyti*, PRT.3 *láidė*, etc. The sound change (*éi* > *ái* after *l*), which affected also Daukša's dialect, makes it impossible to distinguish between PRT.3 *léidė* and PRT.3 *láidė*, which should both appear as ⟨łaide⟩. However, that some or all *ė*-preterite forms printed as ⟨łaide⟩ etc. in Daukša belong to *láidyti* is improbable for the following two reasons. Firstly, other inflectional forms of *láidyti* (i.e. its infinitive, supine, imperative, permissive and conditional moods as well as the present and future tenses with their respective participles) do not occur in Daukša's writings at all. Given the amount of text contained in the *Postilla*, this is hardly accidental. Secondly, in the *Postilla* Daukša used both PRT.3 forms ⟨łaide⟩ and ⟨łaido⟩ equally for both the non-iterative and the iterative preterites of the Polish original. Cf., for instance, Pol. *odpuścił* 'forgave (debt, sins)' translated ⟨atłáide⟩ in 477_{32} and ⟨atłáido⟩ in 361_1 vs. Pol *odpuszczał* 'used to forgive, forgave repeatedly (debt, sins)' translated ⟨atłáide⟩ in 349_{22} and ⟨atłáido⟩ in 346_{41}. This strongly suggests that Daukša did not use *láidyti*.

Table 3. Distribution of preterite forms of *léisti* in the Postilla catholicka

	ė-preterite (*N*=17)			*o*-preterite (*N*=113)		
	SIMPLEX	PREVERB	TOTAL	SIMPLEX	PREVERB	TOTAL
PRT.1SG		5	5			0
PRT.1PL			0			0
PRT.2PL			0			0
PRT.3	1	11	12	9	104	113
TOTAL	**1**	**16**	**17**	**9**	**104**	**113**

the *ė*-preterite (13%). This shows that the innovative preterite is preferred in both cases, suggesting that this asymmetry did at least not prevent the preterite tense forms hosting preverbs from changing into the *o*-preterite.

The next source to be examined is the New Testament translated probably from a Polish original by a member of the Lithuanian Calvinist clergy S. Bogusław Chylinski (appr. 1633–appr. 1665), which was completed in 1658. According to available sources (Kot 1958), Chylinski was a native of Šventežeris near Alytus, who received his education probably in Kėdainiai where Daukša had been born more than a century earlier. The language of Chylinski's translation is, by and large, western Highland Lithuanian of the Kėdainiai region. In Chylinski's text 98 finite preterite tense forms of *léisti* are attested. Their distribution is given in Table 4.[13]

Table 4. Distribution of preterite forms of *léisti* in Chylinski's New Testament

	ė-preterite (*N*=51)			*o*-preterite (*N*=44)		
	SIMPLEX	PREVERB	TOTAL	SIMPLEX	PREVERB	TOTAL
PRT.1SG		2	2			0
PRT.1PL	7	1	8	1		1
PRT.2PL		1	1			0
PRT.3	7	33	40	6	37	43
TOTAL	**14**	**37**	**51**	**7**	**37**	**44**

13. The data have been taken from the index by Kudzinowski (1964). Each attestation has been checked with the reprint of the original text in Kudzinowski and Otrębski (1958). In Chylinski's text, *léisti* is also attested nearly exclusively with *ai* in the root (printed mostly ⟨ay⟩, rarely ⟨aÿ⟩ or ⟨aj⟩). However, like in Daukša's writings, the *ė*-preterite forms can hardly belong to the iterative *láidyti* whose other forms do not occur in the text at all. This lack of *láidyti* in the language of two late-16th–mid-17th c. writers associated with Kėdainiai is probably not coincidental.

In Chylinski's text, the inherited *ė*-preterite is attested more frequently than in Daukša's writings. Without the ambiguous PRT.2SG, the figures are 51 *ė*-preterite attestations against 44 occurrences of the *o*-preterite. It is again remarkable that the PRT.1SG is attested only in the *ė*-preterite. Taken together, preterite forms which are never exactly repeated in polar answers (i.e. PRT.1SG and PRT.2PL) display 3 attestations of the inherited *ė*-preterite against zero of the recent *o*-preterite. By contrast, in the PRT.3 which is usually repeated exactly in answers, the ratio is 40 *ė*-preterites against 43 *o*-preterites. This indicates that in Chylinski's Lithuanian the change was heading toward a mixed paradigm similar to that implied by Daukša's writings and, accordingly, is consistent with a possible effect of the repeat-induced person asymmetry. The PRT.1PL has an interesting status, as first plural forms may be repeated exactly in the case of an inclusive interpretation, or be replaced by a second plural form in the case of an exclusive interpretation.

As for the asymmetry between verbs with and without preverbs, simplex preterite tense forms are more frequent in Chylinski than in Daukša. Out of all 21 simplex forms, 7 (33%) pertain to the innovative *o*-preterite, whereas this is true for 37 out of all 74 preverb forms (50%). This suggests a potentially higher rate of change for verbs with preverbs than for simplex verbs. This is the opposite of the theoretical expectation.

The same mixed paradigm of *léisti* is documented for the relevant dialect at the relevant time in the first grammar of Lithuanian written in Latin by Daniel Klein, published in print in 1653 and photomechanically reproduced in Balčikonis and Larinas (1957: 67–272). Its author was born 1609 in Kraupischken (district Ragnit, Duchy of Prussia), studied until 1636 in Königsberg and subsequently worked as a Lutheran pastor in Tilsit until his death in 1666 (Krollmann 1967: 338). The language described in the book is western Highland Lithuanian of Prussia, i.e. closely related to, though not identical with, the Kėdainiai variety of western Highland Lithuanian. Inflectional forms of *léisti* are repeatedly mentioned in the grammar, its PRT.1SG is given as *ė*-preterite *at=léidžiau* (printed ⟨atleidźiau⟩ on pp. 88, 89 and ⟨atleidźau⟩ on p. 96). P. 99 contains the following remark:

> "Nota. E duplici conſona Characteriſticæ loco in themate poſita, altera, qvæ in tertia Præsentis recurrit, hîc in Præterito prorſus reſpuitur, ut: *atleidźau* (& *atleidau*) remiſi, in ſec. & tert. perf. *atleidei* / *atleida*. Sic *geidźau* cupivi, *geidei* / *geide* / *ßwenćiáu* ſanctificavi, *ßwenteí*/*ßwente*."[14]

14. "Note. From the double consonant characteristically placed in the theme, the second [consonant], that reoccurs in the third of the present, this is entirely ejected in the preterite, such as: *atleidźau* (& *atleidau*) forgave, in the sec. & th. pers. *atleidei* / *atleida*. The same *geidźau* wished, *geidei* / *geide* / *ßwenćiáu* sanctified, *ßwenteí*/*ßwente*."

It follows that Klein used the more conservative *ė*-preterite PRT.1SG *at=léidžiau* beside the innovative *o*-preterite PRT.3 *at=léido* (printed ⟨atleida⟩).[15] However, he also knew the *o*-preterite PRT.1SG *at=léidau* as an admissible variant form. In this Klein spectacularly confirms the earlier transition from *ė*- to *o*-preterite of *léisti* in the third person than in the first person singular implied by Daukša's and Chylinski's writings for a further local variety of Highland Lithuanian at the relevant time.

To summarize, in the late-16th–mid-17th c. in the western Highland Lithuanian dialect of Kėdainiai the verb *léisti* was in a process of transition from the inherited *ė*-preterite to a more recent *o*-preterite. The corpus data show that two speakers of said dialect, Mikalojus Daukša and S. Bogusław Chylinski, preferred the more recent *o*-preterite of *léisti* in that part of its paradigm which is regularly repeated exactly in repetitional responses, the PRT.3. At the same time, both writers preferred the inherited *ė*-preterite for those paradigmatic forms which are not repeated exactly in polar answers, notably the PRT.1SG. This distribution of the new and the inherited preterite of *léisti* over the paradigmatic forms in the transition period is consistent with the theoretical expectation according to which forms not exactly repeated in polar answers should lag behind in inflectional change. The other asymmetry to be expected in western Highland Lithuanian during the transition from the inherited to the innovative preterite in *léisti*, i.e. the contrast between forms with and without preverbs, finds no support in the data. While this may suggest that there is no effect of conversational priming, this needs to be tested further against larger datasets, as the verb *léisti* without preverbs is relatively rare in both Daukša and Chylinsky. Beside this, it cannot be excluded that the strategy of answering a polar question by repeating only the preverb did not exist in their sub-dialect at the relevant time.

The findings regarding the person asymmetry appear even more significant if the frequency relations are taken into consideration. In line with previous findings on other languages, in spoken Modern Lithuanian the PRT.3 appears to be, at least in most verbs, considerably more frequent than the PRT.1SG. The figures given by the Corpus of Contemporary Lithuanian Language (subcorpus "colloquial speech") are, specifically for *léisti*, 18 occurrences of PRT.3 *léido* against 5 of PRT.1SG *léidau*; for *at=eĩti* 'to come' one finds 87 attestations of PRT.3 *at=ẽjo* against 29 of PRT.1SG *at=ẽjau*. It follows that the transition from the inherited *ė*-preterite to the new *o*-preterite should have affected the more frequent PRT.3 of *léisti* not earlier but, on the contrary, later than the corresponding PRT.1SG. The fact that the data show the opposite implies in our view the presence of an additional factor which must have been powerful enough to overturn the conserving role of fre-

15. The ⟨-a⟩ instead of expected ⟨-o⟩ in PRT.3 ⟨atleida⟩ is recurrent in Klein's grammar and seems to reflect the pronunciation in his native Ragnit area local dialect (Buch 1998: 103–104).

quency. We submit that this additional factor was the effect of priming through repetitional responses. Of course, more case studies are needed in order to lend firm support to our hypothesis. In this section, we sketched out what such case studies could look like.

6. Conclusion and outlook

In this paper, we set out to sketch possible ways of investigating language change at the enchronic scale of real-time conversations. Specifically, we proposed that conversational priming in repetitional responses may be a conversational mechanism propagating the spread of innovations among individuals and across populations. We conducted two case studies laying some ground work for further investigations of the phenomenon. First, we showed that languages differ regarding the frequency with which they employ repetitional responses. This finding yields further directions for future research: Do innovative forms spread faster across a community which employs repeats with a high frequency? We furthermore argued that even in languages where repeats are employed less frequently, their occurrence rate is high enough to show an effect on the spread of linguistic innovations.

Second, we argued that certain asymmetries introduced by repeats may become visible in the relative speed of inflectional change. We showed that in Lithuanian, the preterite tense of the verb *léisti* 'to let' changed first in the third person and was only later followed by the first person singular, which is in line with the expectations we formed on the basis of the person asymmetry. Of course, we need many more case studies of this kind in order to corroborate the effect of the repeat-induced person asymmetry. If it can be shown that in languages where the person asymmetry holds, innovations spread faster in the third than in first and second person, that gives us one part of the evidence. However, we also need evidence from languages where the person asymmetry does not exist. In these languages, it is predicted that the speed of change should not be higher for the third person. One next step for investigating the potential effect of conversational priming in repetitional responses on language change is thus to gather these two types of evidence in a systematic way. Multi-genre language documentation corpora as envisioned by Nikolaus Himmelmann will undoubtedly form one of the cornerstones of future empirical investigations of language change at the enchronic scale.

Funding

This research received funding from the University of Cologne Excellent Research Support Program, funding line FORUM, project *Conversational priming in language change*, as well as funding line Cluster Development Program, project *Language Challenges*. The Yurakaré data used in the first case study were collected with the help of funding from the Volkswagen Foundation (grant numbers 81821 and 83448).

Acknowledgements

We dedicate this work to our colleague Nikolaus Himmelmann, whose unfaltering dedication to generating truly novel knowledge has been a constant source of inspiration. For many inspiring comments, we are truly grateful to the two anonymous reviewers of our paper as well as to Uta Reinöhl. We would also like to express our gratitude to all the people who participated in the data collection of the cited corpora. All remaining errors are ours.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following glosses are used:

A	set A (specific to Yucatec Maya)	PART	particle
APP	apperceptive	PN	proper name
B	set B (specific to Yucatec Maya)	POL	polite
CL	clitic	PP	preterite participle
COM	committal	PRT	preterite
EGO	egophoric	PST	past
EXP	expletive	REA	realis
HON	honorific	REL	relativizer
INT	intimate speech level	REP	reportative
INTERJ	interjection	TRANS	transferentive

References

Aldrup, Marit. 2024. "Asking the Obvious: Other-Repeats as Requests for Reconfirmation." *Contrastive Pragmatics* 5 (1–2): 274–306.

Aleksandravičius, Juozas. 2011. *Kretingos tarmės žodynas*. [Dictionary of the Kretinga dialect]. Vilnius: Lietuvių kalbos institutas.

Ambrazas, Saulius, Gražina Belickienė, Elena Grinaveckienė, Aldona Jonaitytė, Jonina Lipskienė, Kazys Morkūnas, Birutė Vanagienė, and Aloyzas Vidùgiris. 1991. *Lietuvių kalbos atlasas III. Morfologija*. [*Atlas of Lithuanian language. Vol. III. Morphology*.] (LKA III). Vilnius: Mokslas.

Ameka, Felix K., and Marina Terkourafi. 2019. "What if…? Imagining Non-Western Perspectives on Pragmatic Theory and Practice." *Journal of Pragmatics* 145: 72–82.

Auer, Peter, and Frans Hinskens. 2005. "The Role of Interpersonal Accommodation in a Theory of Language Change." In *Dialect Change: Convergence and Divergence in European Languages*, ed. by Peter Auer, Frans Hinskens, and Paul Kerswill, 335–357. Cambridge: Cambridge University Press.

Balčikonis, Juozas et al. (eds.). 1957. *Pirmoji lietuvių kalbos gramatika 1653 metai*. Vilnius: Valstybinė politinės ir mokslinės literatūros leidykla.

Baumann, Andreas, and Lotte Sommerer. 2018. "Linguistic Diversification as a Long-Term Effect of Asymmetric Priming: An Adaptive-Dynamics Approach." *Language Dynamics and Change* 8 (2): 253–296.

Bauer, Anastasia, Sonja Gipper, Tobias-Alexander Herrmann, and Simona Sbranna. This volume. "The Multimodal Turn in Corpus Building: Lessons from Language Documentation and Sign Language Corpora." In *The Documentarist Turn: From Observable Linguistic Behaviour to Typological Generalizations*, ed. by Sonja Riesberg, Uta Reinöhl, and Birgit Hellwig. Amsterdam, Philadelphia: John Benjamins.

Blythe, Richard A., and William Croft. 2012. "S-Curves and the Mechanisms of Propagation in Language Change." *Language* 88 (2): 269–304.

Bock, J. Kathryn. 1986. "Syntactic Persistence in Language Production." *Cognitive Psychology* 18 (3): 355–387.

Bock, Kathryn., Gary S. Dell, Franklin Chang, and Kristine H. Onishi. 2007. "Persistent Structural Priming from Language Comprehension to Language Production." *Cognition* 104 (3): 437–458.

Brown, H. Paul, Brian D. Joseph, and Rex E. Wallace. 2009. "Questions and Answers." In *New Perspectives on Historical Latin Syntax: Volume 1. Syntax of the Sentence*, ed. by Philip Baldi, and Pierluigi Cuzzolin, 489–530. Berlin/New York: Mouton de Gruyter.

Brown, Penelope, Mark A. Sicoli, and Olivier Le Guen. 2021. "Cross-Speaker Repetition and Epistemic Stance in Tzeltal, Yucatec, and Zapotec Conversations." *Journal of Pragmatics* 183: 256–272.

Buch, Tamara. 1998. *Opuscula Lithuanica. Wydał Wojciech Smoczyński.* [Papers on Lithuanian. Ed. by Wojciech Smoczyński.] Warszawa: Uniwersytet Warszawski.

Chia, Katherine, and Michael P. Kaschak. 2022. "Structural Priming in Question–Answer Dialogues." *Psychonomic Bulletin and Review* 29 (1): 262–267.

Dekker, Peter, Sonja Gipper, and Bart de Boer. 2024. "Conversational Priming in Repetitional Responses as a Mechanism in Language Change: Evidence from Agent-Based Modelling." *Linguistics Vanguard* 10 (1): 549–564.

Deppermann, Arnulf, Alexandra Gubina, Katharina König, and Martin Pfeiffer. 2024. "Request for Confirmation Sequences in German." *Open Linguistics* 10 (1): 20240008.

De Smet, Isabeau, and Freek Van de Velde. 2019. "Reassessing the Evolution of West Germanic Preterite Inflection." *Diachronica* 36 (2): 39–180.

Dvořák, Boštjan. 2007. "Slovenske zaimenske naslonke in zakaj so nekaj posebnega. [Slovenian Clitic Pronouns and What Is So Special about Them.]" *Slovenski jezik — Slovene Linguistic Studies* 6: 209–233.

Eggeling, Julius. 1882–1900. *The Satapatha-Brâhmana: According to the Text of the Mâdhyandina School. 5 vols.* Oxford: The Clarendon Press.

Endzelin, J. 1909. "Zum Lettischen Präteritum." *Zeitschrift für Vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen* 43: 1–41.

doi Enfield, Nick J. 2010. "Questions and Responses in Lao." *Journal of Pragmatics* 42 (10): 2649–2665.

doi Enfield, Nick J. 2014. *Natural Causes of Language: Frames, Biases, and Cultural Transmission*. Berlin: Language Science Press.

doi Enfield, Nick J. 2025. "The Enchronic Envelope." *Psychological Review*, Advance online publication.

doi Enfield, Nick J., Tanya Stivers, Penelope Brown, Christina Englert, Katariina Harjunpää, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Tiina Keisanen, Mirka Rauniomaa, Chase W. Raymond, Federico Rossano, Kyung-Eun Yoon, Inge Zwisterlood, and Stephen C. Levinson. 2019. "Polar Answers." *Journal of Linguistics* 55 (2): 277–304.

doi Fang, Hongmei. 2025. Echo Answers. *Linguistic Typology* 29 (2): 269–317.

Farrell, Meagan T., Lise Abrams, and Katherine K. White. 2012. "The Role of Priming in Lexical Access and Speech Production." In *Psychology of Priming*, ed. by Nobuaki Hsu, and Zacharias Schütt, 205–244. New York: Nova Science Publishers.

doi Floyd, Simeon. 2018. "Egophoricity and Argument Structure in Cha'palaa." In *Egophoricity*, ed. by Simeon Floyd, Elisabeth Norcliffe, and Lila San Roque, 269–304. Amsterdam/Philadelphia: John Benjamins.

Gerullis, Georg. 1923. *Mosvid. Die ältesten litauischen Sprachdenkmäler bis zum Jahre 1570*. Heidelberg: Winter.

Gijn, Rik van, Vincent Hirtzel, Sonja Gipper, and Jeremías Ballivián Torrico. 2011. *The Yurakaré Archive*. The Language Archive, MPI Nijmegen. https://hdl.handle.net/1839/8df587ed-3d6e-4db8-bfe5-4ecad5cef3a2

doi Gipper, Sonja. 2017. "Pre-Semantic Pragmatics Encoded: A Non-Spatial Account of Yurakaré Demonstratives." *Journal of Pragmatics* 120: 122–143.

doi Gipper, Sonja. 2020. "Repeating Responses as a Conversational Affordance for Linguistic Transmission: Evidence from Yurakaré Conversations." *Studies in Language* 44 (2): 281–326.

doi Gipper, Sonja. 2024. "Request for Confirmation Sequences in Yurakaré." *Open Linguistics* 10 (1): 20240026.

doi Gipper, Sonja, and Jeremías Ballivián Torrico. 2014. *Family Problems Yurakaré*. Items SocCog103 through SocCog-109 at PARADISEC.

doi Gipper, Sonja, Katharina König, and Kathrin Weber. 2024. "Structurally Similar Formats Are Not Functionally Equivalent across Languages: Requests for Reconfirmation in Comparative Perspective." *Contrastive Pragmatics* 5 (1–2): 195–237.

Gippert, Jost, Javier Martinez, Agnes Korn, and Roland Mittmann (eds.). 2016. *TITUS: Thesaurus Indogermanischer Text- und Sprachmaterialien*. https://titus.uni-frankfurt.de/indexd.htm

Hackstein, Olav. 2021. "Lat. *negō** 'not me; ich nicht', 'I deny; ich verneine', Lexikalisierung von Echo-Antworten und delokutive Ableitung." In *Studies in General and Historical Linguistics, offered to Jón Axel Harðarson on the occasion of his 65th birthday*, ed. by Matteo Tarsi, 223–236. Innsbruck: Institut für Sprachwissenschaft.

Hackstein, Olav. 2024. "Responsive Particles in Indo-European from an Etymological Perspective." Paper read at The Speakers of Indo-European and their World. 150 years of Indo-European studies in Basel. XVII. Fachtagung der IG. Basel, 9 September 2024.

Harjunpää, Katariina, and Ana Cristina Ostermann. 2023. "Responding to Polar Questions in Brazilian Portuguese: É-Responses and Repeats." In *Responding to Polar Questions across Languages and Contexts*, ed. by Galina B. Bolden, John Heritage, and Marja-Leena Sorjonen, 76–108. Amsterdam/Philadelphia: John Benjamins.

Hayashi, Makoto. 2010. "An Overview of the Question–Response System in Japanese." *Journal of Pragmatics* 42 (10): 2685–2702.

Hill, Eugen. 2022. "Conversational Priming through Repeating Responses as a Factor in Inflectional Change." In *Paths through Meaning and Form: Festschrift Offered to Klaus von Heusinger on the Occasion of His 60th Birthday*, ed. by Chiara Gianollo, Łukasz Jędrzejowski, and Sofiana I. Lindemann, 101–105. Cologne: University of Cologne USB Monographs.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–196.

Himmelmann, Nikolaus P. 2006. "Language Documentation: What Is It and What Is It Good for?" In *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin/New York: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2016. "What about Typology Is Useful for Language Documentation?" *Linguistic Typology* 20 (3): 473–478.

Hirtzel, Vincent. 2021. "Entre política identitaria y narrativas autobiográficas. Restituciones digitales de un proyecto de documentación lingüítica en Bolivia (Proyecto DoBeS Yurakaré 2006-2011)." *Journal de la Société des américanistes* 107/1.

Hoffmann, Bettina, and Nikolaus P. Himmelmann. 2009. *Münster Videokorpus Alltagsgespräche*. Unpublished corpus of video recordings.

Holmberg, Anders. 2016. *The Syntax of Yes and No*. Oxford: Oxford University Press.

Hoymann, Gertie. 2010. "Questions and Responses in ǂĀkhoe Hai||om." *Journal of Pragmatics* 42 (10): 2726–2740.

INE (Instituto Nacional de Estadística). 2015. *Censo de población y vivienda 2012 Bolivia: Características de la población*. La Paz: Instituto Nacional de Estadística. https://www.ine.gob.bo/index.php/publicaciones/censo-de-poblacion-y-vivienda-2012-caracteristicas-de-la-poblacion/

Jacobs, Cassandra L., Sun-Joo Cho, and Duane G. Watson. 2019. "Self-Priming in Production: Evidence for a Hybrid Model of Syntactic Priming." *Cognitive Science* 43 (7): e12749.

Jäger, Gerhard, and Anette Rosenbach. 2008. "Priming and Unidirectiona Language Change." *Theoretical Linguistics* 34 (2): 85–113.

Janda, Laura A., Tore Nesset, and R. H. Baayen. 2010. "Capturing Correlational Structure in Russian Paradigms: A Case Study in Logistic Mixed-Effects Modeling." *Corpus Linguistics and Linguistic Theory* 6 (1): 29–48.

Jones, Bob Morris. 1999. *The Welsh Answering System*. Berlin/New York: Mouton de Gruyter.

Kaiser, Julia. 2018. Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien. *Gesprächsforschung — Online-Zeitschrift zur verbalen Interaktion* 19: 515–552.

Kaschak, Michael P., Timothy J. Kutta, and Jacqueline M. Coyle. 2014. "Long and Short Term Cumulative Structural Priming Effects." *Language, Cognition and Neuroscience* 29 (6): 728–743.

Keevallik, Leelo. 2010. "Minimal Answers to Yes/No Questions in the Service of Sequence Organization." *Discourse Studies* 12 (3): 283–309.

Kim, So Yeon. 2016. Repetitional Responses in Korean Conversation. PhD thesis, UCLA.

König, Katharina, and Martin Pfeiffer. 2024. "Requesting Confirmation or Reconfirmation across Languages: An Introduction." *Contrastive Pragmatics* 5 (1–2): 1–26.

Kootstra, Gerrit Jan, and Hülya Şahin. 2018. "Crosslinguistic Structural Priming as a Mechanism of Contact-Induced Language Change: Evidence from Papiamento-Dutch Bilinguals in Aruba and The Netherlands." *Language* 94 (4): 902–930.

Kot, Stanisław. 1958. "Chylinski's Lithuanian Bible: Origin and Historical Background." In *Chylinskis Lithuanian Bible. The New Testament. Vol. II: Text*, ed. by Czesław Kudzinowski, and Jan Otrębski, xliii–lxvii. Poznań: Ossolineum in Wrocław.

Krollmann, Christian. (eds.) 1967. *Altpreußische Biographie. Bd. II: Maltitz–Z.* Marburg/Lahn: Elwert.

Kruopas, J., A. Lyberis, J. Palauskas, J. Senkus, B. Tolutienė, et al. 1968–2002. *Lietuvių kalbos žodynas.* [*Dictionary of Lithuanian languages.*] (LKŽ). Vol. I–XX. Vilnius: Mintis.

Krause-Lerche, Anne. 2022. "Conservation in Ongoing Analogical Change: The Measurement and Effect(s) of Token Frequency." *Corpus Linguistics and Linguistic Theory* 18 (1): 77–114.

Kudzinowski, Czesław. 1964. *Chylinski's Lithuanian Bible. The New Testament. Vol. III: Index.* Poznań: Ossolineum in Wrocław.

Kudzinowski, Czesław. 1977. *Indeks-słownik do "Daukšos Postilė". T. I: A–N. T. II: O–Ž.* Poznań: Universytet im. Adama Mickiewicza.

Kudzinowski, Czesław, and Jan Otrebski. 1958. *Chylinski's Lithuanian Bible. The New Testament.* Vol. II: Text. Poznań: Ossolineum in Wroclaw.

Lepner, Theodor. 1690/1744. *Der preusche Littauer oder Vorstellung der Namens-Herleitung, Kind-Tauffen, Leibes- und Gemüths-Beschaffenheit, Kleidung, Wohnung, Nahrung und Acker-Bau, Speise und Tranck, Sprachen, Gottes-Dienst, Begräbnisse und andere dergleichen Sachen der Littauer in Preussen.* Danzig: Rüdiger.

Levelt, Willem J. M., and Stephanie Kelter. 1982. "Surface Form and Memory in Question Answering." *Cognitive Psychology* 14 (1): 78–106.

Lühr, Rosemarie. 2018. "Prosody in Indo-European Corpora." In *Farnah: Indo-Iranian and Indo-European Studies in Honor of Sasha Lubotsky*, ed. by Lucien van Beek, Alwin Kloekhorst, Guus Kroonen, Michaël Peyrot, Tijmen Pronk, and Michile de Vaan, 173–190. Ann Arbor/New York: Beech Stave Press.

Mahowald, Kyle, Ariel James, Richard Futrell, and Edward Gibson. 2016. "A Meta-Analysis of Syntactic Priming in Language Production." *Journal of Memory and Language* 91: 5–27.

Mair, Christian. 2017. "From Priming and Processing to Frequency Effects and Grammaticalization? Contracted Semi-Modals in Present-Day English." In *The Changing English Language: Psycholinguistic Perspectives*, ed. by Marianne Hundt, Sandra Mollin, and Simone E. Pfenninger, 191–212. Cambridge: Cambridge University Press.

Marijuán, Pedro C., Raquel del Moral, Sungchul Ji, Marta Gil Lacruz, Juan David Gómez-Quintero, and Jorge Navarro. 2019. "Fundamental, Quantitative Traits of the 'Sociotype.'" *Biosystems* 180: 79–87.

Martins, Ana Maria. 2013. "Emphatic Polarity in European Portuguese and Beyond." *Lingua* 128: 95–123.

Miyaoka, Osahito. 2012. *A Grammar of Central Alaskan Yupik (CAY)*. Berlin: Mouton de Gruyter.

Morris, Richard. 1876. *Cursor Mundi (The cursur o the world). A Northumbrian poem of the XIVth century*. Ludgate Hill: N. Trübner.

Nilsson, Jenny. 2015. "Dialect Accommodation in Interaction: Explaining Dialect Change and Stability." *Language and Communication* 41: 6–16.

Ostermeyer, Gottfried. 1791. *Neue Littauische Grammatik*. Königsberg: G. L. Hartung.

Palionis, Jonas. 2000. *Mikalojus Daukšos 1599 metų Postilė ir jos šaltiniai*. [*The 1599 Postil by Mikalojus Daukša and its sources*.] Vilnius: Baltos lankos.

Pickering, Martin J., and Simon Garrod. 2017. "Priming and Language Change." In *The Changing English Language: Psycholinguistic Perspectives*, ed. by Marianne Hundt, Sandra Mollin, and Simone E. Pfenninger, 173–190. Cambridge: Cambridge University Press.

Plaza Martínez, Pedro, Carlos Callapa, Alexander Frontanilla, Daniel Guzmán, Libertad Pinto, Mayra Ponce, Guido Machaca, Ronald Bruno, Mónica Yabeta, Miguel Núñez, Gladys Nogales, and Claudina Parada. 2011. *Historia, lengua, cultura y educación de la Nación Yurakaré*. Cochabamba: FUNPROEIB Andes/CEPY.

Potočnik, Tomaž. 2023. "Saying Yes without Yes: The Positive Response System in Latin." *Transactions of the Philological Society* 121 (1): 65–90.

Reineke, Silke, Arnulf Deppermann, and Thomas Schmidt. 2023. "Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch (FOLK): Zum Nutzen eines großen annotierten Korpus gesprochener Sprache für interaktionslinguistische Fragestellungen." In *Korpora in der germanistischen Sprachwissenschaft*, ed. by Arnulf Deppermann, Christian Fandrych, Marc Kupietz, and Thomas Schmidt, 71–102. Berlin/Boston: de Gruyter.

Reķēna, Antoņina. 1998. *Kalupes izloksnes vārdnīca. 1. sējums: A–M. 2. sējums: N–Ž*. [*A dictionary of the Kalupe dialect. Vol. 1: A-M. Vol 2: N-Ž*.] Rīga: Latviešu valodas institūts.

ReN-Team. 2021. *Reference Corpus Middle Low German/Low Rhenish (1200–1650); Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)* (v. 1.1).

Reinöhl, Uta. This volume. "More Diachronic Than You Think: Historical Depth Within Language Documentation Corpora and Its Potential to Mitigate Two Major Biases in Linguistics." In *The Documentarist Turn: From Observable Linguistic Behaviour to Typological Generalizations*, ed. by Sonja Riesberg, Uta Reinöhl, and Birgit Hellwig. Amsterdam, Philadelphia: John Benjamins.

Rosemeyer, Malte, and Scott A. Schwenter. 2019. "Echoic and Non-Echoic Confirming Affirmative Responses in Spoken Brazilian Portuguese." *Journal of Pragmatics* 141: 80–101.

Rossano, Federico. 2010. "Questioning and Responding in Italian." *Journal of Pragmatics* 42: 2756–2771.

Rossi, Giovanni. 2020. "Other-Repetition in Conversation across Languages: Bringing Prosody into Pragmatic Typology." *Language in Society* 49: 495–520.

San Roque, Lila, Simeon Floyd, and Elisabeth Norcliffe. 2018. "Egophoricity: An introduction". In *Egophoricity*, ed. by Simeon Floyd, Elisabeth Norcliffe, and Lila San Roque, 1–78. Amsterdam/Philadelphia: John Benjamins.

Scheibman, Joanne. 2001. "Local Patterns of Subjectivity in Person and Verb Type in American English Conversation." In *Frequency and the Emergence of Linguistic Structure*, ed. by Joan L. Bybee, and Paul J. Hopper, 61–89. Amsterdam: John Benjamins.

Senn, Alfred. 1966. *Handbuch der litauischen Sprache. Band I: Grammatik*. Heidelberg: Winter.

Seržant, Ilja A., and George Moroz. 2022. "Universal Attractors in Language Evolution Provide Evidence for the Kinds of Efficiency Pressures Involved." *Humanities and Social Sciences Communications* 9(1): 58.

Sorjonen, Marja-Leena. 1996. "On Repeats and Responses in Finnish Conversations." In *Interaction and Grammar*, ed. by Elinor Ochs, Emanuel A. Schegloff, and Sandra A. Thompson, 277–327. Cambridge: Cambridge University Press.

Steinhoff, Hans-Hugo. 1995. *Lancelot und Ginover I. Prosalancelot I*. Frankfurt am Main: Deutscher Klassiker Verlag.

Stivers, Tanya. 2010. "An Overview of the Question–Response System in American English Conversation." *Journal of Pragmatics* 42: 2772–2781.

Strautiņa, Marta. 2007. *Mārcienas izloksne*. [*The dialect of Mārciena*.] Rīga: Latviešu valodas institūts.

Su, Danjie. 2016. "Grammar Emerges through Reuse and Modification of Prior Utterances." *Discourse Studies* 18 (3): 330–353.

Subačius, Giedrius, Mariusz Leńczuk, and Wiesław Wydra. 2010. "The Earliest Known Lithuanian Glosses (~1520–1530)." *Archivum Lithuanicum* 12: 31–70.

Thomas, Earl W. 1969. *The Syntax of Spoken Brazilian Portuguese*. Nashville: Vanderbilt University Press.

Trudgill, Peter. 1986. *Dialects in Contact*. Oxford: Blackwell.

Utka, Andrius, Erika Rimkutė, Jolanta Kovalevskaitė, Agnė Bielinskienė, Mažvydas Petkevičius, Rūta Petrauskaitė, and Jurgita Mikelionienė. *Corpus of the Contemporary Lithuanian Language*. Kaunas: Vytauto Didžiojo universitetas.

Weber, Kathrin. 2024. "Request for Confirmation Sequences in Low German." *Open Linguistics* 10 (1): 20240019.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. "ELAN: A Professional Framework for Multimodality Research." *International Conference on Language Resources and Evaluation (LREC)* 5: 1556–1559.

Zellers, Margaret. 2021. "An Overview of Forms, Functions, and Configurations of Backchannels in Ruruuli/Lunyala." *Journal of Pragmatics* 175: 38–52.

Zinkevičius, Zigmas. 1966. *Lietuvių dialektologija. Lyginamoji tarmių fonetika ir morfologija*. [*Lithuanian dialectology. A comparative phonology and morphology of the dialects*.] Vilnius: Mintis.

Zinkevičius, Zigmas. 1977. "M. Mažvydo raštų kalba." *Baltistica* 13(2): 358–371.

Zinkevičius, Zigmas. 1978a. "M. Mažvydo raštų kalba." *Baltistica* 14(1): 38–44.

Zinkevičius, Zigmas. 1978b. "M. Mažvydo raštų kalba." *Baltistica* 14(2): 139–146.

CHAPTER 28

On similarity, mistake, and belief revision

The case of *yimarne(k)* in Kunbarlang

Ivan Kapitonov & Caroline Gentens
University of Cologne | KU Leuven

The marker *yimarne(k)* in Kunbarlang can be used to signal "mistaken belief", whereby a proposition held as a belief at some point is now considered false. The various functions and formal types of expressions *yimarne(k)* combines with point to an analysis in terms of two separate markers, a short form with a core function of marking similarity, and a long form expressing counterfactuality. This suggests that the long form's mistaken belief use might stem from an original similarity marker with an added element of unclear origin. To further assess the potential pathway whereby a similative acquires counterfactual uses including mistaken belief, we compare the case of *yimarne(k)* to attested markers of mistaken belief both within the Gunwinyguan family and beyond.

Keywords: mistaken belief, similarity, multi-perspectival marker, Kunbarlang, Gunwinyguan, semblative, polyfunctionality, counterfactual, grammaticalization

> The point where descriptive approaches diverge most sharply in dealing with multifunctional affixes is with regard to what evidence is considered strong enough for weakening the "one form — one meaning" principle (and the intensity with which such evidence is sought). Of major importance in this regard is the question of what is used to define "one form" in this equation.
>
> (Himmelmann 2006: 496)

1. Introduction

Kunbarlang (Gunwinyguan, Australia) has two particles, *yimarne* and *yimarnek*, with somewhat overlapping but distinct functions. In this paper, we explore these particles as related forms which arguably show a functional extension from the domain of similarity (1): to a kind of focus, on the one hand, and to propositional operators such as mistaken belief[1] or unrealised action (as in (2)), on the other.[2]

https://doi.org/10.1075/slcs.240.28kap
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

(1) *karlu panikin la mankorlng ngadda-marnbuni ngadda-marnbuni*
no dish and carpentaria_palm 1PL.IRP-make.IRP 1PL.IRP-make.IRP
[jiməŋɜ] panikin bonjbonj
Y dish too
'We had no tin and we used to make *mankorlng* [carpentaria palm] like a tin.'
(20060620IB03/04:22)

(2) *[jiməŋɛk] ngayi-kangkayini kabbala la nga-ngunda eh-eh*
YK 1SG.IRP-go.IRP boat and 1SG.NF-do.PST INTJ
'I was like, maybe I'll go on the boat, but then I was like, nope.'
(20060620IB05/02:26)

Both forms appear in Joy Kinslow Harris's thesis (Kinslow Harris 1969) and field notes, the earliest available extensive records, but with only one token each. The first thorough description of these particles is in Carolyn Coleman's unpublished thesis, where they receive clearly distinct treatments. Coleman analyses the shorter of them as a one-of-its-kind predicate:

> *yimarna* is a restricted-concord predicator which is the only member of its word-class. Although it is a two-place predicator it takes no concord affixes at all; it is also strikingly similar in form and meaning to 'seemingly real' status adverb *yimarnek*.
>
> (63) *ngalbuk yimarna durduk*
> she Y dog
> 'she resembles a dog (in looks or behaviour)' [gloss ours — IK and CG]'
>
> *yimarna* can be named an "irrealis existential predicator": it marks irrealis status of a type of 'unreal existence' which it predicates of its subject.
> (Coleman 1982: 133)

The longer form, on the contrary, is analysed as an adverb:

> Peripheral status 'non-positive' operator *yimarnek* is a non-real adverb which occurs in clause-peripheral Adverbial phrases… Informants translate it as "I thought"; it is used as English forms "I thought", "as though", or "it seems as if" are used. *yimarnek* can therefore be described as indicating "seemingly real (but not real)" status…

---

1. Under "mistaken belief" we understand, following McGregor (2024: 106), "a dual modal perspective on a proposition… [representing] it as someone's belief whilst simultaneously asserting its falsity".

2. The cited data come from different collections: the labels that start with C02 indicate data from Coleman (1984); the labels that start with {2006, 2014, 2015} indicate data from O'Keeffe and Singer (2015). Finally, examples whose labels start with IK1 come from Kapitonov (2016).

(58) *yimarnek ngutta girri-bomo*
YK you.SG 2SG-stupid
'It seems as if you're stupid (although I know you're really not)' [gloss ours — IK and CG]

> In clauses which contain verbal predicators the "seemingly real" status is marked both with *yimarnek* and in the morphosyntactic components on the verb…
> (Coleman 1982: 131)

Thus, the two particles are delineated in terms of form and distribution: *yimarnek* is clause-peripheral and can combine both with nominal and with verbal predicates/clauses.[3] In contrast, *yimarna* occurs in the canonical transitive verb position between the subject and the object, and does not combine with verbal predicates/clauses (as its complement). However, they are not very clearly delineated in terms of semantics: both have to do with the category of "status",[4] the longer form indicating the "seemingly real (but not real)" status, whereas the shorter form "marks irrealis status of a type of 'unreal existence'" (Coleman 1982: 131, 133).

In the first author's fieldwork, these distinctions proved to be less categorical. Inclined to view the non-reality contribution of both forms as a functional unity and facing the challenge posed by Kunbarlang's lenition processes and occasional transcription difficulties, he soon stopped paying close attention to the form. In his 2021 grammar, the distinction is, without any discussion, reduced to free variation, which leads to a unified discussion as *yimarne*(*k*), the "similative and quotative particle" (Kapitonov 2021: 325).[5]

The present two authors set out to write about that single polyfunctional particle and its potential evolution from a similative marker to a quotative particle, on the basis of a transcribed corpus of examples. Naturally, the second author, who has not had first-hand experience working with Kunbarlang but has diligently studied the sources, wanted to know what the deal was with the two forms described by Coleman. A more careful retranscription of all tokens — both the first author's own materials and archival ones — followed, and the *yimarne(k)* story has taken a slightly different twist. This paper is still about the semantic development of a marker, but it is also a story about mistaken belief revision, and the role of primary data in setting things right.

---

3. "Status adverbs… rarely occur in the Predicator Phrase; they usually occur at clause peripheries" (Coleman 1982: 109).

4. Coleman's "status" is defined for verb stems and corresponds to our "mood", which is defined for verb stems and personal prefixes in a system of composite TAM marking.

5. Cf. also Kapitonov and Gentens (2018).

In exploring the two Kunbarlang particles we shall face the following questions, some of which will be harder to answer than others. What is the actual distribution of either particle? What is the full range of (attested) functions of either particle? Are the forms connected, and precisely how (and similarly for the functions)? To what extent do the observed patterns resemble what we know about other languages, whether within or outside of Australia?

The examination will reveal that, *pace* Kapitonov (2021), the distinction between the two forms is not categorically clear-cut. However, tendencies observed in our corpus clearly support Coleman's (1982) analysis with two distinct items. Since the connection between them has been pointed out ("*yimarna*… is also strikingly similar in form and meaning to 'seemingly real' status adverb *yimarnek*", Coleman (1982: 131)) but was not discussed any further, we shall undertake making sense of this connection. In doing so we shall also draw on the cognate particles in the closely-related Bininj Kunwok, and then on other Australian languages where similar polyfunctionality is found.

2. Primary data

In this section we first give a brief introduction to Kunbarlang, so that the readers can follow the examples. Then we present our data set and explain the methodology. After that we turn to the descriptive core of this study, in which we examine the distribution of the two forms over the morphosyntactic contexts and over the functions that they have.

2.1 Kunbarlang

Kunbarlang is a non-Pama-Nyungan Gunwinyguan language, spoken in the northern part of Arnhem Land (the "Top End"). It is highly endangered, spoken by an aging population of about 30 individuals, its transmission severely limited, if not completely interrupted.

Like all Gunwinyguan languages, Kunbarlang is polysynthetic with holophrasis, i.e. a verbal wordform can (and in narratives often does) constitute the whole clause. Verbs in their templatic structure index up to two participants, inflect for tense and mood (reality status), and have a number of derivational affixes (applicatives, adverbial meanings, and incorporated nominals). Person indexes cumulatively encode person and number of a participant as well as a mood and a tense value. Person and tense/mood inflection is obligatory, i.e. there are no non-finite forms. Of relevance to the present discussion is the distinction between realis and irrealis mood. The latter is used in the context of negation and for a

range of special functions, such as past habituals, protasis of a conditional, and some other modal uses. The mood system interacts with tense, with a three-way contrast in the realis and only a past/non-past distinction in the irrealis.

Nominals, by contrast, are morphologically austere, noun class agreement on modifiers and a few derivational suffixes being all that is attested. Nominative-accusative alignment is evident from the indexing patterns in verbs. Case morphology exists only in personal pronouns, which distinguish direct (S=A=P), genitive, and dative forms.

2.2 Data set and annotation

Since this study is partly driven by a certain dual perspective, as it were, or a conflict between the "clear-cut" two-form analysis and the "messy-realisation" free-variants analysis, we find it imperative to describe the data and method in detail. Our corpus consists of 84 examples from recordings spanning nearly 40 years from 1982 to 2018 and made in Warruwi, Maningrida, and Darwin by Carolyn Coleman, Isabel O'Keeffe, Ruth Singer, Aung Si, and Ivan Kapitonov. Of these 84 examples, 23 come from conventional elicitation, the rest from narratives, conversation, and picture description tasks. Each token was marked for text type, so as to be able to restrict the analysis to non-elicitation examples, for instance. Some additional tokens beyond those 84 were excluded from the analysis — these are mainly repetitions or ungrammatical examples from elicitation and dictionaries.

In terms of realisation, not all tokens are uniform. Apart from the "canonical" short form realisation [ˈjimənɻɛ], occasionally truncated realisations are perceived, e.g. [ˈiman], [ˈjimәɻ], etc. The final vowel can be heard as [ɜ] or [a]. The final velar stop of the form [ˈjimənɻɛk] can be unreleased [ˈjimənɻɛk̚]. For this study, we decided to take the surface form at face value, rather than assume that some short form tokens are truncated long forms. We have divided the forms into one of two categories: the short form without acoustic evidence of the final [k] and the long form where [k] is detectable. This way one can explore whether such a surface form grouping yields meaningfully distinct distributions over contexts and/or functions. In what follows we shall operate with the terms *long* and *short form*. In citing examples, however, we give a broad IPA transcription of these two particles (whereas the rest is in the practical orthography). Interestingly, the long forms are only half as frequent in the dataset: 28 long vs. 56 short. Excluding the data from elicitation, the proportion becomes even more uneven: 11 long vs. 50 short.

In addition to the short/long form, the data were coded for the following features. Based on the characterisation of the two particles by Coleman (1982), we annotated their syntactic scope. Though not exceptionless, the rule is that both particles immediately precede their scope; in addition to the linear order we

were guided by semantics of the utterance and by intonational phrasing. Next, we annotated the morphosyntactic status of the scope, i.e. whether it is nominal or verbal and in the latter case, realis or irrealis. We discuss the distribution of the forms over the morphosyntactic contexts in Section 3.1.

A note on coding: in a polysynthetic, flexible word order language like Kunbarlang it can be hard to decide what an operator scopes over (e.g. just a nominal or the whole clause containing it). To achieve consistency and rely on less arbitrary decisions, we used two main heuristics. First, we chose the smallest object consistent with the given translation. Second, when in doubt about the scope, we relied on intonational phrasing. The contrast between (3) and (4) illustrates this; curly brackets indicate intonational phrases in these and selected other examples.

(3) *{[jimənɛ] banambi bonjbonj} {ngadda-ka}*
Y dragging_net too 1PL.NF-go.NP
'We would go like [with] a dragging net.' (20060620IB04/00:58)

(4) *{[jimənɛ] mandabul kidda-madjingi} {nukka namalurru kadda-rnak-bum kadda-kudjiweng}*
Y barrel 3PL.IRP-pierce.IRP he axe 3PL.NF-just-hit.PST 3PL.NF-??
'It was as if they shot with a shotgun, this axe, they would just hit'em and [?] [i.e. that many geese came down]' (20060620IB04/05:21)

The linear structure in (3) and (4) is very similar: *yimarne* is followed by an instrument noun, which is followed by a verb. The annotation, however, is different. The former datapoint is annotated as a combination with a nominal, since the verb is excluded from the intonation phrase containing *yimarne*. In the latter example, both the phrasing and the translation suggest a larger object than the noun alone, and thus it is annotated as a combination with a clause, its mood value irrealis.

Since we are interested in the meanings expressed by these particles, we annotated them as follows. The two authors independently annotated the function of the particle informally, and then discussed their judgements and together synthesised these into a set of distinct functions. For these we could moreover draw on earlier descriptions like Evans's discussion of the cognate marker in Bininj Kunwok (2003: 609–612, 255–256), on descriptions of unrelated but functionally similar particles in Kayardild (Evans 1995: 378–382) and in Kuuk Thaayorre (Gaby 2017: 329–332), as well as on literature on the grammaticalisation of semblatives (among others Meehan 1991). Finally, we grouped the resulting 15 functions into seven "macro-functions" in order to make our generalisations, including quantitative ones, more robust. In what follows, we refer to those macro-functions, but we will also discuss which ones are more, or less, homogenous. These functions and their distribution will be detailed in Section 3.2.

3. Attested distribution: *Yimarnek* and *yimarne*

3.1 Distribution: Scope and mood

In order to explore what syntactic objects the two forms combine with, we annotated the data set with a distinction of four levels: noun, pronoun, clause, and predicate. The latter two essentially reflect our heuristic of choosing the smallest scope. Consider (5) for illustration.

(5) *ngudda benbe [jimәŋɛk] ki-nganj-kangkayini ngayi*
you.SG yesterday YK 2SG.IRP-hither-go.IRP I
ngay-ngun-wuni bilmu ki-djarri
1SG.IRP-2SG.OBJ-give.IRP barramundi 2SG.IRP-eat.IRP
'If you came yesterday, I would've given you barramundi to eat.'
(IK1-170609_1SY2/10:43–54)

The particle is located in the protasis of a counterfactual conditional, but rather than being located at the edge of the clause, it precedes the verb alone. This is the kind of case that we annotate as "scope:predicate".

The scopal feature comprises two contrasting clusters: nouns and pronouns on the one hand, and predicates and clauses on the other. As a separate parameter, we annotated whether the morphosyntactic form of the scope is nominal, realis, or irrealis. This parameter lumps the nominal cluster into a single category, and uses a different feature to cross-cut the verbal cluster. The noun/verb distinction came out conspicuously neat, with only one case (from elicitation) where a clause is scoped over but its predicate is nominal (thus outside of the realis/irrealis system).

Now we can turn to the distribution of the forms within our 84 datapoints: consider Figure 1.[6]

In Coleman's (1982) analysis, the long form modifies clauses, which can have verbal or nominal predicates, whereas the short form is a predicator, itself taking a (nominal) complement. The distribution in Figure 1 violates the expectations based on that analysis and reveals that the asymmetry between the two forms is more attenuated. The long form conforms to the expectations better with its clear preference for the verbal ones ($n = 25$ out of 28). Once elicitation is excluded, the preference becomes even stronger with ten examples out of eleven showing *yimarnek* scope over a verb or a clause. The one example in (6) that formally shows nominal scope has *yimarnek* introduce a highly elliptical utterance, and thus rather fits the pattern. This example is discussed in greater detail below.

6. The plots have been made with the ggplot2 R library (Wickham 2016).

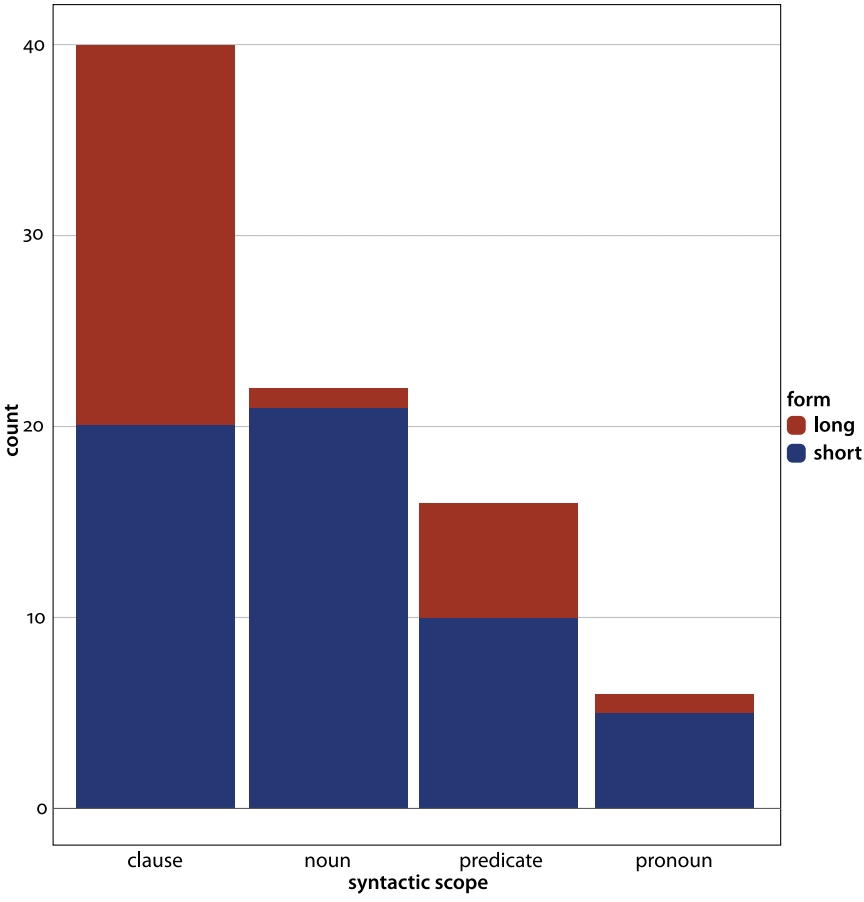


Figure 1. Distribution of forms over syntactic scope

(6) *{"mammam ka-nganj-kanj kikkakin"} {[jimənɛk] karrang}*
MF 3SG.NF-hither-take.NP meat YK mother
{nakarlyung barda nukka[?]}
crocodile what he
'"Grandpa bringing meat!" **Like, [take it for your] mother**. This crocodile or whatsit.' (20140704IOv01/03:09)

The short form in our data takes scope over both nominal ($n=26$) and verbal ($n=30$) objects, however. Noticeable is that it combines with a verbal object more than half of the time. If elicited examples are excluded, the ratio is exactly half, with 25 examples each. There are two possibilities. The less interesting one is that some of the short tokens are actually lenited long tokens, i.e. the target coda containing the velar stop is not pronounced. The other possibility is that most of the

short form tokens are genuine and this form is not confined to nominal objects. We return to this below when the discussion can be informed by the functional distribution of the forms as well.

Turning to mood, we find further discrepancy with what might be expected. Based on Coleman's (1982: 131) observations that the "'seemingly real' status is marked both with *yimarnek* and in the morphosyntactic components on the verb", one might expect a prevalence of irrealis forms in the scope of *yimarnek*. However, the distribution is even, as Figure 2 shows.

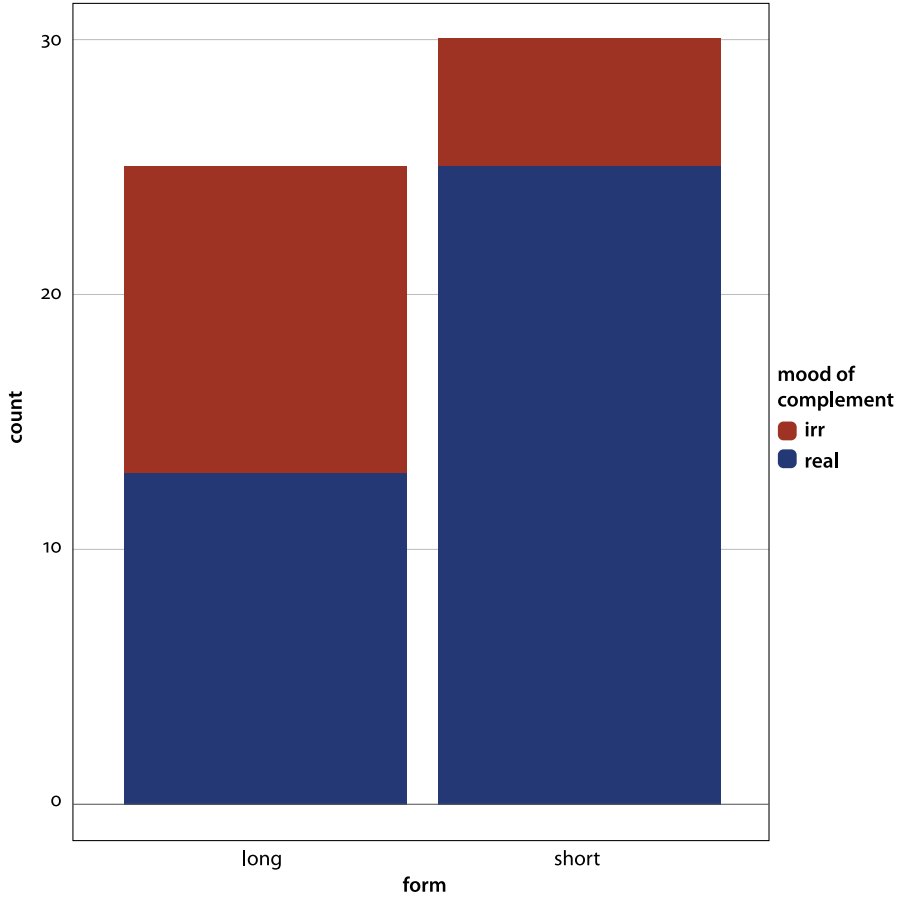


Figure 2. Irrealis and realis in the scope of long and short forms

That is, in the scope of *yimarnek* there are 12 irrealis verbal forms against the 13 realis ones, meaning that the morphosyntactic makeup of the verb is independent of the presence of *yimarnek*. The short form also scopes over realis and irrealis alike. Here, indeed, realis prevails with 25 tokens out of 30.

Examination of examples reveals that the distribution of mood forms reflects the distinction between the actual state of affairs (even if mistakenly construed) on the one hand, and counterfactual situations and future-in-the-past, on the other. In other words, irrealis marks events that either are counterfactual with respect to the speech time (5), or lie in the future with respect to the original belief held. Consider (7).

(7) *{[jimaŋɛk] ngana-kudji* ***ngana-djarri}*** *{but balanda ngemek*
YK 1DU-one 1DU.IRP-eat.IRP but whitefella yet
kabarra-djarrang bonjbonj!}
3DU.NF-eat.PST too
'[We thought] like only we would eat that, but the whitefellas ate it too!'
(20150413IOv01/03:49–53)

If the reported event had taken place prior to the described original belief, then it is marked with realis, even when reported as a mistaken description. Consider the case of belief revision in (8).

(8) *{[jimaŋɛk] djaway* ***ka-bum*** *la nukka ngorro kibuke nayi Wakadj*
YK dingo 3SG.PST-hit.PST and he DEM.MED.IV before ART.I W
ka-maddjing ka-rdam}
3SG.NF-pierce.PST 3SG.NF-put.PST
'They thought a dingo had got it, but it was one that Wakadj had shot before.'
(C02-026607v1/15:24)

This is in line with our current understanding of the functions of the Kunbarlang tense/mood forms: the counterfactual (foreclosed) situations are normally in the irrealis past (cf. Kapitonov 2021: 212). Thus, the use of the irrealis is not conditioned by *yimarnek* per se, but is chosen compositionally based on the semantics of the form.

3.2 Distribution: Range of attested functions

As explained above, we have identified seven "macro-functions" in the data, to which we turn now. For clarity, we capitalise the names of the functions throughout. For the *short form*, the functions are: Comparison ("like/as if it were N"); Focus; Exemplification; Mistaken Belief; Accord Clause; Reportative; and one example of Unrealised Action (see also the summary in Table 1 below).

Among these, there is a striking asymmetry between the most frequent function, Comparison (1, 3), observed with 33 tokens, and the other six functions, over which the remaining 23 tokens are divided. By "comparison" we mean comparison

of similarity (not inequality), where similarity may obtain along various dimensions, including perceptual (9), functional (3), or similarity of circumstance (10).

(9) *{Like} {[**jimaŋa**] ka-yuwa=rdurdu} {[jimaŋa] balmad bonj} {kuyi*
ENG Y 3SG.NF-lie.NP=thunder Y rain exactly ART.IV
ka-buddu-bakbelbunj ngayi kirdimarrk} {nayi mankurdel}
3SG.NF-3PL.OBJ-throw_down.NP ART.PL man ART.I giant
'**It's like when the thunder roars**, when it's wet season, that's how [it sounds like when] he throws the people down; the giant.' (20060901IB08/00:18)

(10) *[jimaŋɛ] ngondo ka-ngunjdje kanak la ngadda-bukayinj*
Y DEM.PROX.IV 3SG.NF-do.NP sun and 1PL.IRP-rise.PST
kenda
DEM.PROX.LOC
'It was like this time when we got here.' (20140709IOv05/06:28–30)

Notice that in this Comparison use the scope of *yimarne* may be either verbal ($n=16$; as in (9) and (10)) or nominal ($n=17$; as in (3)).

Second in frequency, seven further tokens instantiate the macro-function Exemplification, e.g.:

(11) *[jimaŋɛ] Djudda nukka ngorro ka-yuwa na-rleng*
Y Dj he DEM.MED.IV 3SG.NF-lie.NP I-much
'[context: Sometimes leeches bite all of us, they bite us when we go] Like at Djudda there's a big mob there.' (C02-026607v1/01:36)

The next most frequent macro-function of the short form is Focus ($n=6$). This term is used here as in Underhill (1988) and Meehan (1991: 44–46), i.e. close to the information-structural notion of *new information focus*. In discussing the grammaticalisation of English *like*, Meehan (1991: 45) suggests that its function "as focusing on new information begins to be seen in the 'approximately' sense", i.e. when the quantifier in a noun phrase is to be focused upon (cf.: *I wrote it in* like *ten minutes*). The extension continues in the "for example" uses, in which not just the NP but a prepositional phrase or even the entire clause is the focus. In both of these uses, however, *like* "still has enough semantic content that it cannot be removed from the sentence" (ibid.; cf.: *That sounded like a lecture* vs. **That sounded a lecture*). By contrast, further semantic bleaching is manifest in the "focus" sense, in that *like* can be removed with the only apparent effect being a lesser focus on the material following it (cf.: *She* like *covered the mirror*). It seems to be accompanied by syntactic context expansion, in particular the inclusion of predicative adjectives in its scope, as in *…it's like gross*. These criteria of (i) marking new information and (ii) apparent omissibility apply also to a number of Kunbarlang examples, such as (12).

(12) *{maybe Natalie kanj-ka} {korro Wigu} {karnda [jiman] ngorro} {that little island}*
ENG N 3SG.FUT-go.NP DEM.MED.LOC W DEM.DIST.LOC Y DEM.MED.IV ENG
'Maybe Natalie will go to Wigu, like that little island there.'
(20060814IB02/00:49)

One of the weaknesses in applying this category is that at present we cannot ask native speakers for omissibility judgements. Thus, we rely on our own best judgement, which is that the status of *yiman* in (12) is different from (9) above. In the former case, *karnda ngorro*, without the particle, would still mean 'there', even if we cannot comment on the finer difference. In the latter case, *kayuwa=rdurdu* without the particle would mean 'thunder [actually] is roaring', i.e. the meaning of mere likeness would be lost. The final note about this macro-function is that this is where the majority of pronominal scope data fall.

In ten examples the function of *yimarne* has to do with propositional content. Of these, four instances of the short form mark Mistaken Belief (henceforth MB; Breen 1984; McGregor 2024).

(13) *[iman] kadda-ngunda=barr na-bareng, but nukka ngorro karlu, nukka ngorro na-mak, Christian man*
Y 3PL.NF-do.PST=think I-dangerous ENG he DEM.MED.IV no he DEM.MED.IV I-good ENG
'They thought he was dangerous, but he wasn't, he was good, a Christian man.'
(IK1-160624_0001/03:05–12)

Notice that the particle alone is sufficient for belief ascription, thus often there is no matrix thought/speech verb (cf. (7), (8)).

Next, there are three Accord Clauses, which function similar to reported speech/thought, but differ from reported speech constructions in that a generic proposition is reported, rather than a specific speech (or thought) event.

(14) *{[iməɽ] ngarrk-ngundje Nangarridj kun-rnungu yalbi Kuwarnda}*
Y 1IN.NF-say.NP N IV-he.GEN country K
'Like we say that it's Nangarridj's country, Kuwarnda.' (20150413IOv01/04:40)

At least two tokens introduce reported discourse, one of which arguably involves a Reportative with indirect speech (15). The speaker is telling about her conversation with someone and first reports their speech directly, as (i) evident in the use of first person pronoun in reference to the reported speaker and (realis) future for that speaker's plans at the time of reported utterance, and (ii) additionally indicated by a global raising of the pitch register for the duration of the quote (see Bishop (2002: 367–368) on tonal register raise in Bininj Kunwok). Then she returns to the

normal pitch range and utters *yimarne* and third person reference to that reported agent. The slashes / … \ mark the extent of the high pitch excursion.

(15) */ nganj-ka […] nganj-buddu-wardam […] \ [jimənɛ] ki-rdolkkandji*
1SG.FUT-go.NP 1SG.FUT-3PL.OBJ-ask.NP Y 3SG.IRP-awake.IRP
'"*I'll go [there to see them], I'll ask them [if they are cooking what they stole]*."
Like, as if he was awake.' (C02-026607v1/17:11–19)

Finally, there is one more example with the meaning of Unrealised Action ("UA", more specifically, a vain attempt; see below).

(16) *{[jimənɛ] ngarra-bun-kala la} {karlu}*
Y 1SG.IRNP-3SG.OBJ-get.IRP and no
'I tried to grab it [a little horse] but nothing.' (20060814IB01/03:14)

The attested functions of the *long form* are a subset of those of the short form: Unrealised Action ($n = 12/28$); Mistaken Belief; Reportative; Focus; and one instance of Exemplification. The Unrealised Action macro-function includes such specific uses as unrealised intentions (17), frustrated wishes, vain attempts, averted events, and counterfactual conditionals.[7]

(17) *{[jimənɛk] ngayi-kangkayini kabbala la nga-ngunda eh-eh}*
YK 1SG.IRP-go.IRP boat and 1SG.NF-do.PST INTJ
'I was going to go on the boat, but I said [to myself] "nope"'.
(20060620IB05/02:26)

In line with our observation above that the counterfactual situations are encoded by the past irrealis form of the verb, the verb is in the irrealis mood in 10 out of the 12 examples of this use.

For the MB uses of *yimarnek*, see (7) and (8). There are eight such data points, two of them elicited.

The macro-function Reportative ($n = 5$) is mainly attested in elicited examples, plus one narrative example (18) that did not have a translation and is open to interpretation (repeated from (6) above).

(18) *{"mammam ka-nganj-kanj kikkakin"} {[jimənɛk] karrang}*
MF 3SG.NF-hither-take.NP meat YK mother
{nakarlyung barda nukka[?]}
crocodile what he
'"*Grandpa bringing meat!*" Like, [take it for your] mother. This crocodile or whatsit.' (20140704IOv01/03:09)

7. Cf. the term *abortive* in Ngakulmungan Kangka Leman (1997: 197).

In this narrative, the speaker, FA, is bringing goannas after hunting. First, he shifts into a high pitch intonational contour (as in (15) above) as he quotes someone back at the camp announcing his return: "*Mammam kanganjkanj kikkakin*!" — 'Mother's father is bringing meat!'. Then back in his normal range he says: "*Yimarnek karrang*" — 'Y mother'. The previous annotation had no translation of this passage, but only the comment from the speaker's grandchild LN (more specifically, the daughter's child, who he would reciprocally call *mammam*): "he told me, take it for your mother; yimarnek karrang". The most plausible interpretation here is this: in the described episode LN, seeing FA return, announces his return. This is quoted by FA, marked by the high plateau. FA tells LN to take the catch to someone who is the speaker's daughter and the child's mother (hence: *karrang*). This is quoted or reported using *yimarnek*, but not pitch.

Two further examples are classified as Focus, e.g. (19).

(19) *{nayi [jimaŋɛk]} {kaddum wohka ka-yuwa}* *{mayi kulkurdu}*
ART.I YK top part 3SG.NF-lie.NP ART.III rope
'[there is] a rope that is, like, partly on top [of the table]'
(IK1-170610_2SM1/28:14)

The one example of the Exemplification function (20) comes from elicitation and is interesting insofar as it forms a near-minimal pair with a short form.

(20) *kunmak [jimaŋɛk] kuyi ngundji-nganj-ka=way*
good Y ART.IV 2DU.FUT-hither-go.NP=hither
'It's okay for you to come.' (IK1-180516_1PN2/46:36)

This form is produced right after a short one. Another difference is that only the *yimarnek* instance is used together with the article *kuyi*, whereas *yimarne* is used on its own. There is no impression of the second token manifesting repair. The lesson that we take from this pair is: the two forms can in principle be used in the same function/context in rapid succession.

Now that all the macro-functions of the two particles have been illustrated, we can turn to the summary of their distribution, given in Table 1 and plotted in Figure 3.

As we consider the full scope of the functions found with the two forms, a number of asymmetries are evident. First, Comparison is exclusively served by the short form. Accord Clauses appear to be similarly restricted to the short form, but there are too few tokens to make a strong generalisation. Next, within the long forms, there is clear prevalence of Unrealised Action and MB over the other functions. Furthermore, these two main uses of the long form appear to be less important for the short form (MB $n=4$) or outright marginal (UA $n=1$). Indeed, the five functions with overlap all show skewing towards one of the forms. The long

Table 1. Summary of the attested macro-functions

	Short	Long
Compar	33	NA
UA	1	12
MB	4	8
Exempl	7	1
Focus	6	2
Report	2	5
Accord	3	NA

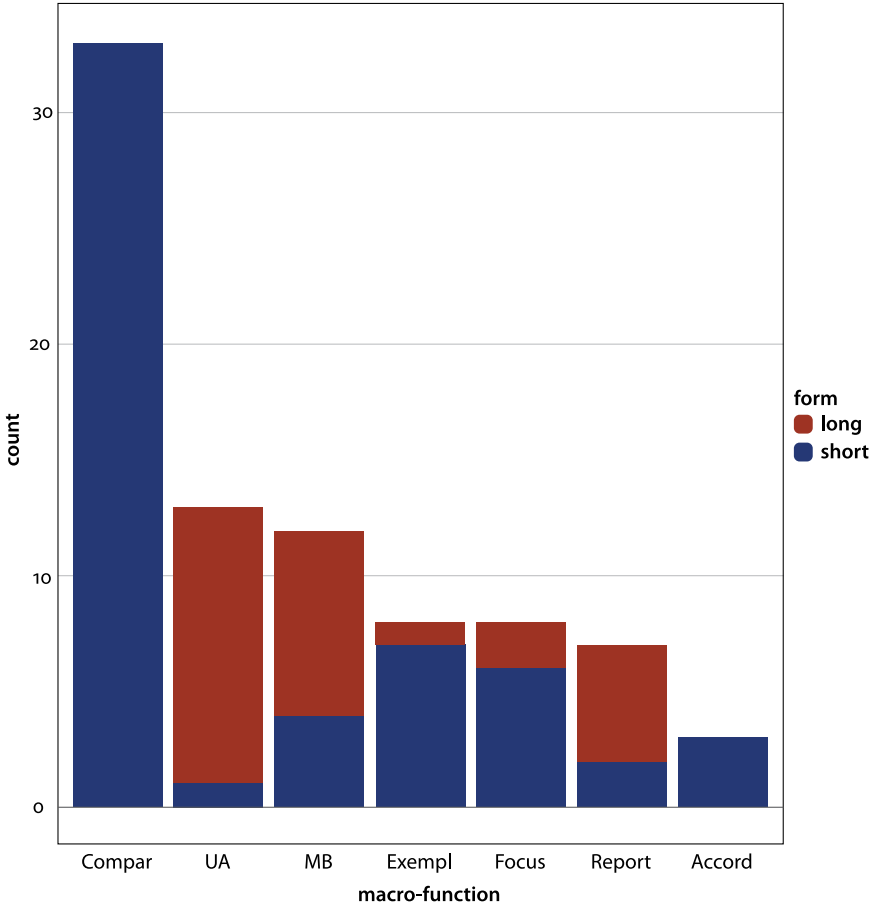


Figure 3. Distribution of the macro-functions

form prevails in the cluster of three functions related to events and propositions: UA, MB, and Report. Focus and Exemplification, on the other hand, are mainly served by the short form.

3.3 Summary

Combining observations from Sections 3.1 and 3.2, we can now confirm that despite some overlap, the quantitative picture favours the analysis with two functionally differentiated forms (*pace* Coleman 1982). They show clearly diverging tendencies. Lenition of the final velar, which we alluded to in the introduction, cannot possibly explain the narrower distribution of the long form: it is conspicuous that we do not find any k-final tokens in the context of comparison of similarity (see Figure 3). It is likewise conspicuous that the k-final forms combine very well with clauses or verbal predicates, but only marginally so with nominals (see Figure 1).

At the same time, it is just as clear that the distinction is not categorical, and that the short form does not have comparable distributional restrictions. It has strong functional tendencies (towards Comparison, Focus, Exemplification), but some tokens are attested in each of the functions where the long form prevails. Correspondingly, it scopes not only over nominals, but over verbal predicates/clauses as well (see Figure 1). Thus, we appreciate that in the face of the "striking similarity in form and meaning" (cf. Coleman 1982: 133), and seeing the points of distributional overlap, one may be led to believe that the two are but free allomorphs (as Kapitonov 2021 did). It is the primary data that enabled us to (i) revise the existing transcriptions, (ii) observe the distributional tendencies, and as a result (iii) correct certain false beliefs.

The similarity relation between *yimarne* and *yimarnek* is clearly not accidental. But what is the relation between them exactly? In the next section we turn to this issue, contrasting our Kunbarlang findings with what is known about other Australian languages.

4. Similarity and mistake in Gunwinyguan and other Australian languages

The polyfunctionality of a form between the "similarity" and "counterfactual" senses (usually with the MB and/or UA flavour) is attested in a number of unrelated and geographically discontinuous languages in Australia. This was first pointed out for two Karnic languages (Pama-Nyungan; Central Australia) by Breen (1984), and later discussed in Kalkatungic and Paman, and non-Pama-Nyungan Tangkic and Gunwinyguan (see the supplementary materials to McGregor (2024) for the list, as well as the discussion below for details).

Harkins (1986: 567), in a discussion of the Warlpiri form *kulanganta* where the sense of 'similarity' is not otherwise mentioned, offers the following observation. She has "sometimes heard Aboriginal speakers of English say *look like…* or *like as if…* in contexts similar to" what we here classify as MB. While admittedly anecdotical, this is nevertheless an interesting piece of evidence. In standard English *like* has not developed this function, and some translations really do not paraphrase easily using *like* — for instance (21), in which the MB marker *kulanganta* is translated with 'I thought'. At the same time, Gentens (2024) has argued that present-day English *as if*, "combining an element of similarity (*as*) with an element indicating hypotheticality" (Kortmann 1997: 318), functions as a mistaken belief marker, on its own as a response particle and when introducing insubordinate clauses (see also Brinton (2014) and López-Couso & Mendez-Naya (2012a, 2012b) on such uses and their diachronic development).

(21) *Kulanganta-lpa yantarli nyinaja, kala lawa.*
KULANGANTHA-IPFV home sit.PST but no
'I thought that he was at home, but he wasn't.' (Warlpiri; Laughren 1982: 149)

Three scenarios are logically possible. First, there may be a more abstract meaning that contextually surfaces as what the (prevailingly English-speaking) researchers identify as either "similarity" or "counterfactual". In McGregor's (2024: 134) words, "the semblative sense is thus not emically distinct from the mistaken belief sense". This meaning is hard for the outsider researchers to put a finger on, but is completely natural for the speakers of those languages where it is lexicalised by a single marker.

Second, there may be an extension from the (arguably more concrete) semblative semantics to the counterfactual uses. Third, there may be the opposite development, i.e. from the counterfactual core to the semblative uses. In our present state of knowledge, it is impossible to have a definitive answer as to which of these scenarios is, or are, the actual case(s). In what follows, we review the evidence for these possibilities and find that in any event the conclusions are not

trivial. We shall also outline what further work can be done to further our understanding of this semantic area.

4.1 Gunwinyguan: Semblative to counterfactual

We first review the cognate particles in Bininj Kunwok, the closest relative of Kunbarlang, and then argue that in both languages evidence points to the development from semblative to counterfactual.

The cognate particles in Bininj Kunwok, likewise, come in two forms, the short and the long. The short form is *yiman* 'like', apparently identical across all dialects. The long form, glossed by Evans as 'counterfactual', varies across the dialects: *(yi)mankek* (Gun-djeihmi), *yimankek~yimanek* (Kunwinjku), *yimarnek* (Kuninjku), and in Kune "a further alternative form *yimanmakken*, apparently synonymous with *yimankek* etc., is also sometimes heard" (Evans 2003: 609). Apart from Kune with its extra-long form, all varieties contain the formative *-(k)ek*. In parallel to Kunbarlang, the short form "most commonly introduces expressions of similarity" (Evans 2003: 255).

(22) *Kodowele na-kimuk-ni yiman Namorrodo na-kimuk-ni-duninjh.*
K M-big-PST Y N M-big-PST-really
'Godowele was big like Namorrodo, who was very big.'
(Kunwinjku; Evans 2003: 121)

Besides Comparison of similarity (22), Evans (2003: 255–256) identifies the short Bininj Kunwok form *yiman* as having the Exemplification function. Examination of further examples throughout the grammar suggests to us still other uses. First, the Focus use is well-attested (p. 245 ex. 6.50; p. 247 ex. 6.58; p. 315 ex. 7.246, 7.250; p. 524 ex. 11.176; or p. 688 ex. (14) where a listener initiates repair using it). Furthermore, there is one Accord Clause (p. 709 ex. (3), one MB (23), and one Unrealised Action use (p. 687–688 ex. (13)). We have not found examples only of the Reportative use of *yiman*.

(23) *Yiman birri-kodj-yime-ng na-wern rerrih bininj, but nungkah*
Y 3AUG.P-head-do-PPFV M-many behind person but he
na-djal-kudjih~kudji ø-ngudj-bebme-ng.
M-just-RDP~one 3P-speed-emerge-PPFV
'Everyone thought that there were lots of people there, but there was just him on his own doing it all.' (Kune; Evans 2003: 355)

The long form "is used to represent a descriptor nominal, or the state of affairs in a clause, as someone's belief… [which] is almost always false" (Evans 2003: 610–611).

That is, it is also functionally very similar to the Kunbarlang long form, its dominant use being MB and UA (including uses like 'nearly', 'almost').[8]

(24) *Bindi-na-ng bedda la yimanek kunj, wanjh bindi-yame-ng.*
3AUG/3PL-see-PPFV them and YK kangaroo then 3AUG/3PL-spear-PPFV
'They saw them and thought they were kangaroos (failing to realise they were really dangerous snake beings), then speared them.' (Kuninjku; Evans 2003: 610)

In these two Gunwinyguan languages a very similar picture is at hand. There are two related forms. The longer form has the narrower distribution (a strong quantitative preference for combining with clausal objects) and is mainly associated with "counterfactual" readings. The shorter one has "similarity" as its most prominent function, but encroaches on the "counterfactual" functions as well. This is clearer in Kunbarlang, perhaps partly due to the difference in the data we looked at (since our Bininj Kunwok "data set" contains just the examples pre-selected for the grammar).

This picture is best explained as follows (using the Kunbarlang forms for illustration). There is a marker *yimarne* with the original meaning of comparison of similarity. The similarity is concrete, i.e. perceptual, and corresponds to the nominal scope. It undergoes semantic evolution, gradually developing uses that are further removed from its original function. One of the directions is towards the MB and UA uses. This semantic development includes two principal aspects (cf. McGregor 2024: 135). With regards to form, the scope expands to include clausal elements.[9] With regards to meaning, there is strengthening of the implicature: what is similar is not identical (cf. Breen 1984: 3; Smith & Johnson 2000: 437). Together these yield the "counterfactual" sense. In some of these extended uses a longer form *yimarnek* (presumably, from *yimarne* + another gram X; see below on its possible identity) is developed, and subsequently becomes specialised. This in turn leads to a reduction in the short form's occurrence in those contexts — say, our MB and UA macro-functions — a possible case of reduction in semantic-pragmatic usage contexts for the short form by "curtailing" under competition with the newly grammaticalising long form (cf. Himmelmann 2005: 85). The long form *yimanmakken* in the Kune variety can additionally support the morphological analysis with alternative formatives extending the original short form: *-makken* in Kune, *-(k)ek* in other Bininj Kunwok varieties and in Kunbarlang.

8. Evans's own preference is to think about this as belief revision, rather than mistaken belief (see e.g. 2003: 611).

9. The fact that dedicated similarity markers can have variable scope may run against this speculation (Gentens & Boye 2024; Wiemer 2024).

In the absence of sufficient historical data, only this general speculation is possible, not stepwise observations about context expansion. Another outstanding question is the identity of the "gram X". We return to these issues briefly at the end of the paper. For now, we turn to reviewing the evidence for the opposite development.

4.2 Karnic and Tangkic: Counterfactual to semblative?

McGregor (2024: 134) outlines the pathway that could lead in the opposite direction. He observes that in mistaken belief constructions, the belief is sometimes not attributed to anyone in particular. What is often enough conveyed is that "the proposition was such that it could be or could have been universally held to be true, i.e. by anyone" (Laughren 1982: 150). Such contexts can serve as bridging from "the coded meaning 'it was mistakenly believed'" (McGregor 2024: 134) to a contextual reading/conversational implicature that the "mistaken observation [may be] a result of similarity", in Breen's (1984: 8) words.

Breen's account of the enclitics =*nyadi* in Yandruwandha and =*widi* in Pitta-Pitta (both Karnic languages of Central Australia) was the first paper about the affinity of "similarity" and "mistake" in Australian languages.[10] Indeed, Breen (1984: 1) suggests that their "primary meaning [may be] related to the idea of mistake (potential or actual) rather than resemblance". This may suggest, but of course does not have to entail an opposite diachronic direction to our proposal about Gunwinyguan.

These two particles display a range of uses very similar to the ones we have been discussing for Kunbarlang. Breen's (1984: 8) analysis, brief as it is, provides an insight about a different possible connection between the meanings. He identifies the two main features of meaning as "similarity" and "deception", the latter including three subtypes: (i) deliberate deception, or pretence, (ii) self-deception, and (iii) deception caused by (a change of) circumstances. Consider, for example, an "unfulfilled intention" in (25).

(25) *Thatjilinga**widi** ngantja, nhuwaka karntakayangu.*
eat-ing-for-like I he-there go-did-hither
'That fellow came along just as I was going to have a feed.'
(Pitta-Pitta; Breen 1984: 5)

10. See also Breen (2004: 231–233) for more Yandruwandha examples. We reproduce Pitta-pitta and Yandruwandha examples as given in the source, and only boldface the relevant morphemes, =*widi* and =*nyadi*.

The link between "deception" and "similarity" is found in examples where the "'like' phrase is used as a description of behaviour, but there is also the statement or implication that observers would be deceived" (Breen 1984: 8).

Presumably this inference can then conventionalise to yield purely semblative uses, such as (26):

(26) *Putiyita**nyadi** thayingatji.*
potato-like eat-next-emph
'They ate them like potatoes.' (Yandruwandha; Breen 1984: 2)

Breen (1984) draws a distinction between "similarity without identity" (*He looks like Bill*) and "similarity with possible identity" (*It looks like Bill*; see the Pitta-Pitta marker *=widi* used in this way in (27)). The latter crucially rules out the similarity-as-basis-for-mistake reading. We conjecture that the availability of such readings for a given marker is evidence for the "similarity" being coded and not merely implied.

(27) *Thithi**widi** nhuwaka tharupaliya.*
brother-like he-there speak-now
'It sounds like my brother.' (Pitta-Pitta; Breen 1984: 3)

Particularly instructive for reasoning about this pathway are data from languages where the "counterfactual" gram does *not* encode comparison of similarity, such as the Warlpiri particle *kulanganta* (28):

(28) *Kulanganta-ka nyiya yangka kurlarda-piya wirnpirlinya-yani. Ngula-nya*
KULANGANTA-IPFV what ANAPH spear-like whistle-INF-go-NP ANAPH-FOC
waripirapira-ju.
whip_snake-DEL
'You think it's something whistling past like a spear. That's a whip snake.'
(Warlpiri; Laughren 1982: 149, glossing original)

This example is relevant to the present discussion in at least three aspects. First, note the dedicated semblative morpheme *-piya*. Second, the meaning here is not strictly that of a mistake: it seems true that something *is* whistling past *like a spear*. Rather, this is a prime example of the interpretation that "the proposition was such that it could be or could have been universally held to be true, i.e. by anyone" (Laughren 1982: 150), but at the same time speaker suggests that "anyone would think it is not true if they knew what I have come to know", as Harkins's (1986: 567) Natural Semantic Metalanguage-style explication puts it. Here the speaker knows better, as they have constructed that image as an illustration of what a whip snake('s behaviour) is like. Third, here the similarity sense seems to be relevant, inasmuch as encounters with a whip snake *resemble* perceiving something whistle past like a spear.

So far, we have seen two patterns. Kunbarlang and Bininj Kunwok feature two forms, the shorter associated with "similarity", the longer with "counterfactual". Accordingly, we offered an account in which the semantic development reflects the formal containment: "similarity" → "counterfactual". In both Pitta-Pitta and Yandruwandha there is a single form for both senses. This is most directly compatible with the state of affairs where the two senses are contextually appropriate manifestations of a more general underlying semantics.[11] However, an argument has been made to the effect of "counterfactual" → "similarity". The Tangkic languages Kayardild and Lardil offer the missing pattern, where the longer form has similarity uses, whereas the shorter one seemingly does not.

Kayardild has the particle *maraka* with a similar range of meanings (Evans 1995: 378–381). On the one hand, there is a family of counterfactual senses, close to our Unrealised Actions but perhaps with more accent on the discrepancy with the prescribed course of action, and to our Mistaken Beliefs. The particular shades of meaning arise depending on the construction, in particular the verb inflection and modal case in verbal clauses (with potential vs. actual inflection having the same effect as the Kunbarlang irrealis vs. realis). With nominal scope, it seems that the modal locative case favours MB readings, while the nominative case favours Comparison as in (29).

(29) *dan-da kunawuna jungarr*, ***maraka*** *niwan-da kanthathu mankarr*
that-NOM child[NOM] big[NOM] MARAKA his-NOM father[NOM] sturdy[NOM]
'That boy is big, he's sturdy like his father.' (Kayardild; Evans 1995: 381)

In Lardil there are two forms, *mara* and *maraka* (Ngakulmungan Kangka Leman 1997: 197–198; the dictionary seems to suggest that the formative *=ka* is the focus marker, cf. p. 133). The short one seems to be mainly used in counterfactual conditionals and UA. The long form *maraka* has the same counterfactual sense, "but can also appear with non-future predicates, indicating an appearance contrary to fact".

(30) *kiin nguthaliya-ku*, ***maraka*** *kalka-n*
this[NOM] pretend-instantiated MARAKA sick-NMLZ
'He's fooling, (acting) as if he were sick (but he's not really sick).'
(Lardil; Ngakulmungan Kangka Leman 1997: 198, segmentation from Evans 1995: 382)

11. Other such single forms are attested in Arabana-Wangkangurru (Karnic; *=wili*, probably cognate to Pitta-Pitta; Hercus 1994: 136, 254), Yalarnnga (Kalkatungic; *-nhangka*; Breen & Blake 2007: 67), and the two Paman languages Kugu Nganhcara (*ka'i*) and Kuuk Thaayorre (*kar*; Gaby 2017: 328–332). Only the last one is discussed in detail, although Gaby's taking "the most basic of these to be likening one (head) NP to another (modifying) NP" seems offered without argument.

It seems that a semblative use is not attested either in this source or in Klokeid (1976). That is, here (in contrast to Kunbarlang and Bininj Kunwok) only the longer form in Kayardild has attested semblative uses. These (very limited and probably incomplete!) data are, at face value, most compatible with the scenario in which the semantic development corresponding to the formal containment is: "counterfactual" → "similarity".

4.3 Taking stock and looking ahead

We started this section by considering three scenarios, one static (one general meaning) and two diachronic ones with the opposite directionality. An adequate understanding of these developments would be interesting in and of itself, but it could also inform the theory of grammaticalisation. We have not arrived at a firm conclusion, which of these is (or are) correct, yet the results are interesting nevertheless.

It is remarkable that quite a number of unrelated languages of Australia show this polysemy, which as of yet is poorly understood and described in wider typology.[12] The *Lexicon of Grammaticalization* (Heine & Kuteva 2002) lists only complementizers and quotatives as attested targets for SIMILE as *source* (Heine & Kuteva 2002: 273–274), and HOW, MANNER, and SAY as sources for SIMILE as *target* (Heine & Kuteva 2002: 335). Consider the diachronic scenarios. If we take the formally motivated analyses of the functional development seriously (i.e. Gunwinyguan vs. Tangkic), we get a situation of bidirectionality. Note that this does not in itself raise a challenge for the unidirectionality hypothesis, at least not on the view that grammaticalisation consists in context expansion for the grammaticalising construction (Himmelmann 2005). On this view, the unidirectionality amounts to absence of spontaneous context reduction, i.e. one that occurs without the interference of another, competing construction. Applied to the semantic domain at hand, a possible counterexample would be if a gram were used in e.g. Comparison and MB functions at some stage, and at a later stage suddenly only in one of those functions. To the best of our current knowledge, none of the discussed patterns suggest such a development.

Nevertheless, the opposite directions of development in the present context would be relatively noteworthy. One might try to reduce those two cases to one direction (does not matter which). In either event, this would entail an analysis in which the development of one functionally distinct form proceeds by reduction of

12. However, see Olguín Martínez (2021), who finds that hypothetical manner constructions, which overlap with the counterfactual domain in Australian languages, involve a similarity marker in 39 languages of the 61 in his sample, and across all continents.

material. Note that this is not mere regular erosion: in the case at hand it would be selective, preferentially or even exclusively eroding the tokens used in a subset of functions.[13] Such "erosive morphology" appears highly unusual and perhaps quite implausible, yet we do not want to rule this out a priori.

In a way, the static scenario may be the most economical. The formal divergence and functional specification would stand as a challenge for it, but perhaps a lesser one than the challenges outlined for the other options above. This is not an easy matter to decide, yet not impossible, we believe. We concur with McGregor's dictum that "neither meaning, mistaken belief nor semblative, can be presumed synchronically more fundamental or diachronically prior. Both are equally plausible candidates, and require an evidential basis" (McGregor 2024: 135). What further evidential basis do we need then?

The two lines of work that seem the most promising to us are: corpus-based study of primary data, and careful reconstruction. For the former one, we hope to have given convincing illustration in our study of Kunbarlang. The more detailed descriptions (e.g. Bininj Kunwok or Kayardild) explain how the construction including the marker has to be taken into account to yield the different readings. Without wishing to cast doubt on the accuracy of those analyses, we think that the Kunbarlang example teaches us that even a small-scale quantitative study can attenuate a good description (s.a. Coleman 1982) in interesting ways. Besides, in the case of Bininj Kunwok, a closer look than the scope of a grammar can afford has resulted in identifying some more functions of *yiman* than the grammar had acknowledged, which is critical for a more precise analysis of e.g. the relation between the short and the long forms.

The historical line of work is just as necessary. It may not be sufficient for a solid analysis of Kunbarlang and Bininj Kunwok forms to consider a corpus of naturalistic data, however large, at the time-depth available for Australian languages. It is very important to understand what the gram X, which hypothetically extended *yimarne* to *yimarnek*, can plausibly be. If it turns out that the Lardil (and Kayardild) *maraka* do indeed contain the focus marker *=ka*, this would rule out the (admittedly somewhat bizarre) analysis *maraka* 'semblative AND counterfactual' → *mara* 'counterfactual'. Similarly for the Gunwinyguan forms; but we do not know what the Gunwinyguan *-k* / *-ek* / *-kek* can plausibly be. Nick Evans (p.c.) suggests the interjection *kek* 'oh no' as a candidate (https://njamed.com/entry/kek

13. Uta Reinöhl (p.c.) points out that grams can split over different constructions and develop independently, opening the possibility for phonological erosion in a subset of contexts or functions. This analysis is not possible for Kunbarlang, where the short form has a wider distribution both in terms of constructions and semantic functions (see Figures 1 and 3), but it remains a possibility to be explored in a closer, ideally corpus-based, study of the Tangkic forms.

/?spelling=defaultandtab=Dictionary). We think other possibilities, e.g. the interjection *nuk* that introduces alternatives or uncertainty (e.g. Coleman 2010: 101) can be just as convincing. Gaby (2017) entertains the relation of *kar* in Kuuk Thaayorre to the negator *kaar*. Inspired by this, one could speculate that the negator *marrek*, attested in all varieties of Bininj Kunwok as well as in Kunbarlang, may be the historical source. Clearly, speculation could lead one far afield.[14]

Another complication is worth a brief mention. In the Maningrida language Burarra there is a particle *jimarna* (occasionally *jimarn*) "mistaken inference, thwarted intention" (Green 1987: 13). The dictionary elaborates as follows: "This word often shows that the speaker thought one thing, but realised that something else was true. Another meaning is that the speaker isn't really sure about something being true. This word often occurs with other modal words…" (Carew & Glasgow 2024). Examples in Green (1987), Glasgow (1988), and Guerrero Beltrán (2024) all align with what we call MB and UA here, and there are no instances of semblative-like use.

(31) *jimarn jarra wangarra a-na-yurthci-nga apu-la.*
JIMARN instead ghost 3MIN.AN-towards-run-RLS 1MIN.OBL-DAT
'I thought it was a ghost running towards me.' (Burarra; Green 1987: 67)

This could be borrowing from Gunwinyguan, or common inheritance from Proto-Arnhem (see Green (2003) on this grouping of Gunwinyguan and Maningrida). We cannot get into details here, but either scenario is non-trivial, and would have repercussions for the correct analysis of *yimarne*. It should be obvious that a sound reconstruction is of paramount importance in such cases.

5. Conclusion

On the basis of a small corpus study we have adduced novel evidence for the status of the Kunbarlang forms *yimarne* and *yimarnek*. Based on the overlapping but partly diverging distribution, we analyse them as two distinct forms. The former has a wider distribution, centres on the comparison of similarity sense, and is arguably the original form. The latter is less frequent overall, shows a strong preference for scoping over verbal/clausal syntactic objects (rather than nominals), and functions primarily as a counterfactual propositional operator that

14. Consider also this note from Evans (1995: 393): "Suffixes or clitics in *-ka* with elusive function are widespread in Australia — see for example Blake's descriptions of Kalkatungu (Blake 1979: 95–96) and Yalarnnga (Blake 1971: 22)".

yields such readings as mistaken belief, frustrated or unrealised action, or reporting speech/thought.

In an attempt to specify the relation between these two forms more precisely, we have reviewed markers with similar functions (and polyfunctionality) in a number of other Australian languages. While there are several illuminating language-specific descriptions, as well as a typology of means for expressing mistaken belief, the semantic area that connects mistake and similarity in Australian languages (and other language families) needs more systematic characterisation. The data that we have at the moment are not sufficient for strong conclusions. However, we hope to have advanced the discussion, as well as to have demonstrated the interest and relevance of continuing this direction of research.

Acknowledgements

We would like to thank the editors, in particular Uta Reinöhl, for their advice and assistance, and two anonymous reviewers for their feedback. Our analysis has benefitted from discussions with Nick Evans and with the audience of the Syntax of the World's Languages 8 (Inalco, Paris), where an earlier version of this work was presented. Finally, the structure of the narrative was substantially developed in conversations with Steph Yam, for which we are very grateful. Any remaining shortcomings are our own.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In the data cited from other sources, we unify the labels with the Leipzig standards, unless indicated otherwise. The following additional glosses are used:

AUG	augmented
AN	*an* noun class
ANAPH	anaphoric
DEL	delimiting
IRNP	irrealis non-past
IRP	irrealis past
I	noun class 1
II	noun class 3
IN	inclusive
INTJ	interjection
IV	noun class 4
MED	medial
MF	mother's father
MIN	minimal
NF	non-future
NP	non-past
PPFV	past perfective
RLS	realis
Y	*yimarne*
YK	*yimarnek* (also the long forms in Bininj Kunwok).

References

Bishop, Judith. 2002. Aspects of Intonation and Prosody in Bininj Gun-wok: An Autosegmental-Metrical Analysis. PhD thesis, University of Melbourne.

Blake, Barry. 1971. "Jalanga: an outline morphology." *Papers on the languages of Australian Aboriginals*, 12–27. Canberra: Australian Institute of Aboriginal Studies.

Blake, Barry. 1979. *A Grammar of Kalkatungu*. Canberra: Pacific Linguistics.

Breen, Gavan. 1984. "Similarity and Mistake in Two Australian Languages." *Language in Central Australia* 2: 1–9.

Breen, Gavan. 2004. *Innamincka Talk: A Grammar of the Innamincka Dialect of Yandruwandha with Notes on Other Dialects*. Canberra: Pacific Linguistics.

Breen, Gavan, and Barry J. Blake. 2007. *The Grammar of Yalarnnga: A Language of Western Queensland*. Canberra: Pacific Linguistics.

Brinton, Laurel J. 2014. "The Extremes of Insubordination: Exclamatory *as if*!" *Journal of English Linguistics* 42 (2): 93–113.

Carew, Margaret, and Kathleen Glasgow (compilers). 2024. *Burarra and Gun-nartpa Dictionary, Published Online as Gijingarliya Wengga*. Jabiru: Bininj Kunwok Regional Language, Culture Centre. www.janguny.com (Accessed October 2, 2025).

Coleman, Carolyn. 1982. A Grammar of Gunbarlang: With Special Reference to Grammatical Relations. BA Honours thesis, Australian National University.

Coleman, Carolyn. 1984. *Sound Recordings Collected by Carolyn Coleman, 1981–1983*. Archived at AIATSIS (call number COLEMAN_C02) and available at ELAR, handle http://hdl.handle.net/2196/00-0000-0000-000F-BF4E-0

Coleman, Carolyn. 2010. *Kun-Barlang — English Dictionary*. Unpublished Manuscript.

Evans, Nicholas. 1995. *A Grammar of Kayardild: With Historical-Comparative Notes on Tangkic*. Berlin/New York: De Gruyter Mouton.

Evans, Nicholas. 2003. *Bininj Gun-Wok: A Pandialectal Grammar of Mayali, Kunwinjku and Kune*. Canberra: Pacific Linguistics.

Gaby, Alice R. (in collaboration with Kuuk Thaayorre language experts). 2017. *A Grammar of Kuuk Thaayorre*. Berlin/Boston: De Gruyter Mouton.

Gentens, Caroline. 2024. "'As if': The Diachronic Development of a 'Mistaken Belief' Use". Presentation, SLE 2024, Helsinki, August 2024.

Gentens, Caroline, and Kasper Boye. 2024. "Introduction: The Development of Manner Expressions into Complementizers or Quotatives." *Linguistics* 62 (3): 531–539.

Glasgow, Kathleen. 1988. "The Structure and System of Burarra Sentences." *Papers in Australian Linguistics* 17: 205–251.

Green, Rebecca. 1987. A Sketch Grammar of Burarra. BA Honours thesis, Australian National University.

Green, Rebecca. 2003. "Proto Maningrida within Proto Arnhem: Evidence from Verbal Inflectional Suffixes." In *The Non-Pama-Nyungan Languages of Northern Australia: Comparative Studies of the Continent's Most Linguistically Complex Region*, ed. by Nicholas Evans, 369–421. Canberra: Pacific Linguistics.

Guerrero Beltrán, David Felipe. 2024. The Language of Time in Burarra/Gun-nartpa: Interactions of Tense, Aspect, and Modality in a Maningrida Language from Northern Australia. PhD thesis, Université Paris-Cité and University of Melbourne.

Harkins, Jean. 1986. "Semantics and the Language Learner: Warlpiri Particles." *Journal of Pragmatics* 10: 559–573.

Heine, Bernd, and Tania Kuteva. 2002. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Hercus, Luise A. 1994. *A Grammar of the Arabana-Wangkangurru Language, Lake Eyre Basin, South Australia*. Canberra: Pacific Linguistics.

Himmelmann, Nikolaus P. 2005. "Gram, Construction, and Class Formation." In *Wortarten und Grammatikalisierung: Perspektiven in System und Erwerb*, ed. by Clemens Knobloch, and Burkhard Schaeder, 79–93. Berlin/New York: De Gruyter Mouton.

Himmelmann, Nikolaus P. 2006. "How to Miss a Paradigm or Two: Multifunctional *ma-* in Tagalog." In *Catching Language: The Standing Challenge of Grammar Writing*, ed. by Felix K. Ameka, Alan Dench, and Nicholas Evans, 487–526. Berlin/New York: De Gruyter Mouton.

Kapitonov, Ivan. 2016. *Kunbarlang, Central Arnhem Land*. Collection IK1 at https://catalog.paradisec.org.au/collections/IK1" [Open Access].

Kapitonov, Ivan. 2021. *A Grammar of Kunbarlang*. Berlin/Boston: De Gruyter Mouton.

Kapitonov, Ivan, and Caroline Gentens. 2018. "*Yimarne* in Kunbarlang: From Similative to Quotative". Presentation, Syntax of the World's Languages 8, Paris. Available at https://lirias.kuleuven.be/retrieve/516875

Kinslow Harris, Joy. 1969. Descriptive and Comparative Study of the Gunwingguan Languages, Northern Territory. PhD thesis, Australian National University.

Klokeid, Terry Jack. 1976. "Topis in Lardil Grammar." PhD thesis, MIT.

Kortmann, Bernd. 1997. *Adverbial Subordination*. Berlin: De Gruyter Mouton.

Laughren, Mary. 1982. "A Preliminary Description of Propositional Particles in Warlpiri." In *Work Papers of SIL-AAB: Papers in Warlpiri Grammar, In Memory of Lothar Jagst*, ed. by Stephen Swartz, 129–163. Darwin: SIL, Australian Aboriginal Branch.

López-Couso, María-José, and Belén Méndez-Naya. 2012a. "On Comparative Complementizers in English: Evidence from Historical Corpora." In *Creation and Use of English Corpora in Spain*, ed. by Nila Vázquez-González, 309–333. Newcastle upon Tyne: Cambridge Scholars.

López-Couso, María-José, and Belén Méndez-Naya. 2012b. "On the Use of *as if, as though* and *like* in Present-day English Complementation Structures." *Journal of English Linguistics* 40 (2): 172–195.

McGregor, William B. 2024. "On the Expression of Mistaken Beliefs in Australian Languages." *Linguistic Typology* 28 (1): 101–145.

Meehan, Teresa. 1991. "It's Like, 'What's Happening in the Evolution of Like?': A Theory of Grammaticalization." *Kansas Working Papers in Linguistics* 16: 37–51.

Ngakulmungan Kangka Leman. 1997. *Lardil Dictionary*. Mornington Shire Council.

O'Keeffe, Isabel, and Ruth Singer. 2015. *Varieties of Kun-barlang, a Language on the Edge*. ELAR collection available at http://hdl.handle.net/2196/00-0000-0000-0002-EF13-E

Olguín Martínez, Jesús. 2021. "Hypothetical Manner Constructions in World-wide Perspective." *Linguistic Typology at the Crossroads* 1 (1): 6–46.

Smith, Ian, and Steve Johnson. 2000. "Kugu Nganhcara". In *The Handbook of Australian Languages*, vol. 5, ed. by R. M. W. Dixon, and Barry Blake, 357–489. Oxford: Oxford University Press.

Underhill, Robert. 1988. "*Like* Is, Like, Focus." *American Speech* 63 (3): 234–246.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wiemer, Björn. 2024. "Polish *jakoby*: An Exotic Similative-reportive Doughnut? Tracing the Pathway and Conditions of Its Rise." *Linguistics* 62 (3): 729–767.

CHAPTER 29

Grammaticisation of medial speech verbs

A case study from Yali (West-New Guinea)

Katharina Gayler[1] & Sonja Riesberg[1,2]
[1] University of Cologne | [2] CNRS — LaCiTO

Reported speech constructions in many Papuan languages exhibit a range of functions that go beyond the primary function of representing an utterance. For some of these languages, it has been pointed out that a specific verb form — the so called medial-verb form — is required to introduce reported speech in these contexts. This paper presents synchronic evidence for a grammaticisation cline that describes the development from a medial speech introducing clause that introduces reported speech proper in a clause chaining construction to a marker of extended reported speech, a quotative marker, a multi-purpose NP marker, and a marker of epistemic stance.

Keywords: grammaticisation of speech verbs, reported speech, clause chaining, quotative marking, NP marking, epistemic stance

1. Introduction

The grammaticisation of speech verbs and speech introducing clauses (SICs) into grammatical markers of reported speech (*quotatives*) is widely discussed in the literature, and attested for numerous, linguistically diverse languages (see, e.g., Güldemann (2008) on various African languages, Klamer (2000) on Austronesian, Saxena (1988) on Tibeto-Burman, to name just a few). In many Papuan languages, reported speech is omnipresent, and reported speech constructions and their functions in discourse have been studied by a number of authors (e.g. de Vries 1992; Deibler 1971; Healey 1964, Himmelmann & Riesberg to appear; Reesink 1993, 2019; Reesink & Evans to appear; Riesberg to appear a). It has also been observed for various Papuan languages which exhibit clause chaining constructions that it is often the medial verb form of a speech verb, which shows signs of grammaticisation such as semantic bleaching. This has been pointed out most explicitly by Reesink (1993: 223) for Usan, but in some grammatical descriptions we find evidence to suggest that this phenomenon is more widespread in Papuan

https://doi.org/10.1075/slcs.240.29gay
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

languages in general (see, e.g., Berghäll (2015) for Mauwake, Bromley (1981) for Lower Grand Valley Dani, Sarvasy (2017) for Nungon). However, very little is known about how exactly the grammaticisation process of (medial) speech verbs manifests in specific Papuan languages, or cross-linguistically within the different Papuan language families.

In this paper, we investigate the grammaticisation of a medial speech-introducing clause in Yali, a Nuclear Trans-New Guinea language[1] spoken in the Indonesian province Papua Pegunungan, and discuss whether an analysis as a quotative marker is justified. Following Himmelmann (2005), we assume that grammaticisation involves both a grammaticising element and a grammaticising construction. The grammaticising element in our case is the medial verb form *ulug* of the speech verb *u-* 'to say', the grammaticising construction is a medial speech verb that introduces a direct quote and that is part of a clause chain. Two construction types are therefore of crucial importance for this paper: clause chaining constructions and constructions of speech report. We will give a detailed introduction to both in Section 2. Before we do so, the following two sub-sections describe the data we use for our investigation (1.1) and the methods we apply (1.2). In Section 1.3 we give a brief definition of speech introducing clauses and quotative markers. Following the grammatical background in Section 2, Section 3 then describes four different contexts in which we observe features in the use of *ulug* that differ from its core function as a medial verb of saying. In the discussion Section 4 we propose a possible grammaticisation cline for *ulug* and discuss arguments for and against the analysis of *ulug* as a quotative marker. Section 5 briefly summarises our results.

1.1 Data

Our Yali data stems from a language documentation corpus in the sense of Himmelmann (1998), collected between 2012 and 2025 (ongoing):[2] Currently, this corpus comprises more than 11 hours of transcribed, translated and interlinearised recordings of spontaneous and semi-spontaneous speech, including traditional narratives, explanatory texts, personal anecdotes and interviews, retellings, stimulus-based descriptions and dialogic tasks, as well as spontaneous multi-party conversations (see Table 1).

For the discussions in the remainder of this paper, we will exclusively draw on this natural language data and the native-speakers' metalinguistic knowledge pro-

1. We follow the rather conservative classification of Glottolog (Hammarström et al. 2025). See Pawley (2005; 2009) for (re-)assessments of the Trans-New Guinea hypothesis.

2. The corpus is partially available online at The Language Archive (MPI Nijmegen) in the CELD Papua collection (cf. Riesberg et al. 2012–2016) and in Pangloss.

Table 1. Yali corpus of natural texts

Text type	Abbreviation	Amount of data
Conversations	C	01:49:11 hours
Conversational tasks	CT	01:08:01 hours
Descriptions	D	01:12:31 hours
Interviews	I	01:05:34 hours
Narratives	N	01:35:57 hours
Explanatory	E	01:44:10 hours
Retellings	R	00:58:56 hours

vided in their interpretation and translation of this data, rather than on elicited material. In this sense, this paper can also be viewed as a demonstration of Himmelmann's claim that "a corpus of (extensively annotated) primary data documenting linguistic practices and traditions is of use for a variety of purposes" (Himmelmann 1998: 166), including the investigation of grammaticisation processes in a language for which no historical (written) records are available. In the spirit of this volume and the general question that has been a guiding thread in Himmelmann's work, i.e. what does it mean to take spoken language seriously?, we will give a usage-based account of medial speech-introducing clauses in Yali, recognising the crucial role that frequent repetition plays in processes of grammaticisation (cf. Bybee 2003). Furthermore, in working with synchronic natural language data, we follow Himmelmann's (2005) perspective on grammaticisation (see Section 1.2) and focus particularly on the semantic-pragmatic usage contexts of constructions with *ulug* as a key feature of grammaticisation processes.

1.2 Methods

The transcription of our data includes a segmentation into *intonation units* (IUs; in the literature also *intonational phrases* or *intonation phrases*) as a basic unit of spoken language. This segmentation was done by native speakers and the authors in collaboration, and reflects an intuitively perceptible phonetic unit of rhythmic and melodic coherence in the sense of Himmelmann et al. (2018) (cf. also Chafe (1994) for a conceptual discussion of IUs, and Gayler (2025:Chapter 2), for a description of IUs in Yali specifically). An example of this IU segmentation is visualised below as Figure 1 (visualisation obtained using Praat, Boersma & Weenink 2021). We will use the concept of IU as an approximation to prosodic phrasing in this paper.

Concerning semantic interpretations of the constructions discussed, we will rely on the metalinguistic knowledge (cf. Himmelmann 1998) of our Yali research

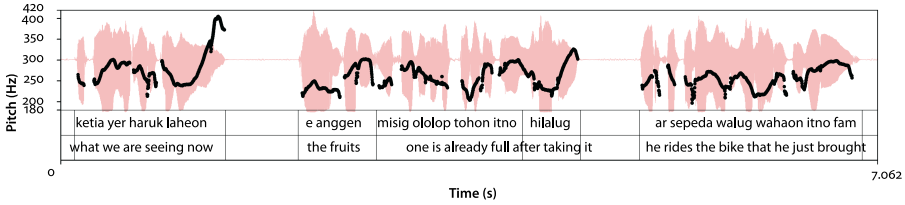


Figure 1. Example of IU segmentation [R pear_story_Lince_029]

collaborators and the choices they made when translating the recordings, as indication of different meanings, for example in the case of possible semantic bleaching.

On a theoretical level, we follow the line of thought in Bybee, who argues that grammaticisation is a process of becoming "automated as a single processing unit" (Bybee 2003: 603), i.e., a transition from a frequent combination into a construction. Joining Reinöhl & Casaretto (2018), we derive a crucial role of synchronic grammatical features and usage-patterns for language change from this argumentation, assuming that certain structures and patterns may favour or hinder such an automatization. For example, we assume that frequent realisations of two adjacent elements as or within one prosodic constituent in natural language use should pave the way for their (re-)interpretation as syntactically, grammatically and/or semantically (more closely) connected elements, e.g., as one lexicalised expression or phrase, or as one syntactic constituent (cf. Himmelmann (2022) on the emergence of grammatical units, also Himmelmann (2014)). Conversely, frequent placement of prosodic boundaries between two adjacent elements should disfavour such a reinterpretation (see Reinöhl & Casaretto (2018), who show this with the non-occurrence of a process of grammaticisation in one branch of Indo-European). Our synchronic language data, we suggest, enable us to show a real-time example of such favourable conditions from within language use.

Furthermore, as noted above, we take a construction-based approach to grammaticisation, and investigate the change of "an element in its constructional context" (Himmelmann 2005: 80). Following Himmelmann (2005), we assume the term 'grammaticisation' to refer to one specific (out of various different possible) type(s) of grammatical change, characterizable as a process of context expansion. It applies to a grammaticising element (*gram*, in our case the medial speech-introducing clause *ulug*) which is in a close local relationship with a host structure (in our case one or more reported clause(s) preceding *ulug*). Both together form the grammaticising construction (here: a medial SIC with a direct quote reporting what someone said), which in itself is again part of a larger syntactic construction (here: a clause chain denoting a sequence of events). Context expansion may occur on three levels: (1) as a host class expansion, where the gram would change semantically to allow more different types of hosts; (2) as a syntactic context expansion, where the whole grammaticising construction would occur in a wider

set of structural environments; and (3) as a semantic-pragmatic context expansion allowing the structure to be used in a broader range of situations. Note that for Himmelmann (2005: 81–83) the third level, the semantic-pragmatic context expansion, is key and a necessary condition of grammaticisation, while the other two are non-obligatory (though most typically, all three levels co-occur). Note as well that processes like phonological erosion of the gram, which are widely associated with grammaticising processes in the literature, though expected to likely occur, are regarded as epiphenomena of the process by Himmelmann. Accordingly, we will not focus on such changes in this paper, but concentrate on semantic-pragmatic usage contexts and syntactic constructions instead.

1.3 Speech introducing clauses and quotatives

In our definition of SICs and quotative markers used for our syntactic discussions in this paper, we will follow Deutscher (2012), who investigates the diachronic development of the latter from the former. SICs, according to Deutscher, are clauses that serve to mark an utterance as a quote, e.g. *and then she said* in (1).[3]

(1) Example SIC (English)
and then she said: 'x'

SICs very typically use verbs of saying, as in (1), but note that under this functional definition, other SICs exist as well, e.g. using the verb 'to go' (2). Although the grammaticising clause in our case is indeed a clause with a verb of saying, this should be kept in mind, as other SICs exist as well in Yali.

(2) Example SIC with *to go* (English)
and then she goes: 'x'

Quotative markers, on the other hand, are grammatical items with non-clausal status that lost (some of the) semantic and syntactic features of a SIC, especially the ability to introduce/denote speakers. At least initially, they retain, however, the function of marking an utterance as a quote. Grammaticising further, this original function may be lost as well, and quotative markers can develop e.g. into complementizers.

---

3. Deutscher (2000: 70) argues that cross-linguistically, the source of the grammaticisation process are indeed full speech introducing clauses, rather than individual elements of SICs, like speech verbs. In Yali, the two cannot always be distinguished: on the one hand, the output structure phonologically preserves indeed only a speech verb in one specific form. On the other hand, however, verbs alone are perfectly acceptable full clauses in Yali. *ulug*, therefore, is a speech verb as well as a possible realisation of a speech clause. Regarding the grammaticisation process, however, we will assume the clause to be the relevant unit.

2. Grammatical background

2.1 Clause chaining

A well-known feature of many Papuan languages is the abundant use of clause chaining structures, i.e. chains of clauses that consist of one or more non-finite clauses, followed by a finite clause as the final 'link' in the chain. In the Papuan literature, the non-finite clauses in these chains are called MEDIAL CLAUSES, with their (non-finite) verbal predicates being referred to as MEDIAL VERBS. As the finite clauses in Papuan clause chains typically occur in final position, Papuanists often refer to them as FINAL CLAUSES, and accordingly the finite verb of such a final clause is sometimes called FINAL VERB. Often, the morphology on the medial verb also indicates the temporal relation between events denoted in the two clauses, e.g. whether they occur simultaneously or sequentially, and whether they share the same subject or not (known as SWITCH-REFERENCE systems). Depending on the language, and often also on the speech genre, Papuan clause chains can be long, consisting of multiple medial clauses, or they can simply consist of one medial verb plus the final verb. In their most prototypical function, clause chains are used to narrate or explain (long or short) sequences of events.

Yali exhibits a number of different clause-linking strategies. Some of these, though sharing some features of chaining, might better be analysed as coordination (see Gayler (2025) for more details), others, however, are clearly clause chaining constructions that exhibit all definitional features mentioned above. Of relevance for the topic of this paper are clause chaining constructions that are formed with the medial-verb marker *-lug* 'same subject sequential (SS.SEQ)'. Medial verb forms with *-lug* express a temporal sequence between events with the same subject. They do not inflect for tense, person and number, and therefore depend on a following, fully inflected verb (the final verb) for their temporal and referential interpretation. An example of a clause chain with *-lug* is given in (3). Both events in the chain (filling potatoes (into a bag) in i., and going down Fasig hill in ii.) happened in the remote past, as marked on the final verb *ambibag*, but in temporal succession.

(3) Clause chain with the medial verb marker *-lug*
suburu itano fani halug Fasig alo ambibag
i. suburu ita=no fani ha-**lug**
sweet.potato DIST=GIV fill 3SG.OBJ-**SS.SEQ**
ii. Fasig alo ambu-pag
Fasig hill go.down-3.RPST
'(i) he put the sweet potato (into his bag), (ii) and went down via Fasig Hill' [N man_and_pig 028]

Long *-lug* chains with several medial clauses are possible as well, but short chains like (3) are much more frequent.

2.2 Reported speech and extended reported speech

In addition to clause chains, the second construction type crucial for the topic of this paper are reported-speech constructions. By far the most frequent type of reported speech in Yali is what has been called DIRECT REPORTED SPEECH in the traditional literature. Direct reported speech takes an 'internal' perspective on the reported event, i.e. pronominal, deictic, or temporal reference is chosen from the viewpoint of the reported speaker.[4] Consider, for example, the data in (4), where both the deictic verbs *ambu-* 'go down' and *waha-* 'come', and the (second person) imperative verb form are chosen from the perspective of the reported speaker (rather than from the perspective of the current speaker). Here and in all subsequent examples, quoted passages in reported speech constructions are indicated by square brackets.

(4) Speech report as direct quote
lek ikma ambik wamag o ehek
[lek ik=mu ambu-ik waha-mag o] ehek
PN river=LOC go.down-DIR come-2.FUT.IMP PTL[5] say:3SG.IMPST
'"come down to the river Lek" he said' [C conversation_2_0851]

Speech introducing clauses in Yali typically contain a verb of saying — in the vast majority of cases a form of *u-* 'say/speak', as in (4)) — but verbs of thinking (especially *pe-* 'think', see (5)) and the generic verb *su-* 'do' are used as well. Both the reported speaker and the reported addressee can be overtly expressed by a noun phrase or a pronoun. If this is the case, they usually precede the quoted passage, as demonstrated in (6) (*he eke tuen* 'this one woman'). Note that, as is the case in (6), NPs denoting the reported speaker are frequently ergative marked, though unlike

4. Yali also exhibits so-called INDIRECT SPEECH, i.e. speech reports which take the 'external' perspective of the current speaker rather than the 'internal' perspective of the reported speaker, as well as BIPERSPECTIVAL SPEECH (cf. Evans 2013: 67), in which the two perspectives are mixed. Both indirect and biperspectival reported speech are extremely rare in Yali, and the current paper will solely discuss instances of direct reported speech (but see Riesberg to appear a, Section 3.5, for examples of the two other types).

5. The particle *o* frequently co-occurs with reported speech in Yali. As discussed in Riesberg (to appear a), it is not clear yet whether it is best analysed as a particle (probably with an interactive function) that is frequently quoted in speech reports, or as a grammaticised or grammaticising marker of quoting itself. The matter being unclear but not under investigation in this paper, we will simply treat it as part of the quote here, aiming for a pragmatic and consistent solution.

in other TNG languages, ergative marking for subjects in SICs is not obligatory (Riesberg to appear a: 7).

(5) Speech introducing clause with *peruk* 'to think'
Yabiahen ano peruk lahi ari
[Yabiahen ano] pe-tuk lat-ihi ari
k.o.tree Q think-INF do-1SG.IMPST MED
'"is it from a Yabiahen tree?" I thought' [C conversation_2_0957]

(6) Realisation of the reported speaker as an NP preceding the quote
he eke tuen ohore nikni Tometten nindi eneg uruken…
he eke tu=en oho=te [n-ikni Tometten n-indi]
woman one PRX=ERG also=TOP 1SG.GEN-father Tometten 1SG.GEN-heart
eneg u-tuk=en
continuously say-INF=CAUS
'because this one woman, too, always says "I miss my father Tometten"…'
[C conversation_2 0040]

The examples above constitute what we might want to call reported speech 'proper', i.e. a speech event that reports what the speaker or another person said (or thought). But it is well known that many languages use reported-speech constructions in a much wider range of contexts, for example, to express purpose, intention, attempt, pretence, apprehension, etc. In the Papuan literature, these functions are often summarised under the label INNER SPEECH (cf. Bromley 1981; de Vries 1990; Reesink 1993), while Spronck and Carsatelli (2021) talk of EXTENDED REPORTED SPEECH. Yali makes extensive use of extended reported speech as well, which is one reason why reported speech constructions are omnipresent in Yali natural discourse. An example is given in (7). In this paper, we suggest that extended reported speech (in Yali) can be viewed as a semantic-pragmatic context expansion of reported speech 'proper', as defined above. We will turn to this point in more detail in Section 3.

(7) Extended reported speech, intentional reading
mene itno / e fene lahamin ulug / pululu hohobo suruk ari
mene itno [e fene laha-min] u-lug pululu hohobo su-tuk ari
dog DET tree on go.up-1SG.IMFUT say-SS.SEQ sound scratch do-INF MED
'the dog wants to climb the tree and is scratching (the trunk)'
LIT: 'the dog says "I will climb the tree" and is scratching (the trunk)'
[R frog_story_Silpa_Fince 099–101]

3. Main uses of *ulug* in Yali

Morphologically, the form *ulug* can be analysed as the sequential medial form of the verb *uruk* 'to say', as in (8), where it functions as a 'simple' medial verb and not as a speech-introducing clause. The glossing as 'say-SS.SEQ' indicates this function. Note that Example (8) does not display a full clause chain, but only two of its medial clauses (for brevity). Here as elsewhere, IU boundaries are represented as slashes (/) in the transcription tier of the examples. Roman numbering indicates a clause segmentation.

(8) Use of *ulug* without reported speech
<ar-> aren oho pasukan kino ulug / wol inaptug
i. at=en oho pasukan **kino u-lug**
3SG=ERG also troop **invite say-SS.SEQ**
ii. wol inap-lug (…)
call 3PL.THEME-SS.SEQ
'he invited the troops, called them, and then (…)'
[E 2nd_initiation_2b 0758–0758]

Uses like (8), however, are comparatively rare and must be actively sought out in our data.[6] Much more typically, *ulug* occurs with reported speech. This is not very surprising, because if the speech verb *uruk* 'to say' can be used to introduce reported speech in a finite clause, as illustrated in Examples (2) and (4) in Section 2.2, it is expected that it is also able to do this in a medial clause of a clause chain. The following two examples contrast these two possibilities. Example (9), once more, shows a final speech introducing clause (preceded by the speech report), Example (10) shows a speech introducing clause in medial position (preceded by the speech report) (in i.), and followed by two other clauses, one medial (in ii.) and one final (in iii.).

(9) *neluk wilahi imisi*
[n-eluk wilat-ihi] **u-misi**
1SG.GEN-alive stay-3SG.IMPST **say-3SG.IMFUT**
'"I was still alive" he will say' [N suit_2 464]

6. Note that *ulug* in Example (8) is part of a light-verb construction, collocating with the coverb item *kino* 'invite' (see Riesberg (to appear b) for a description of light-verb constructions in Yali, and Riesberg and Olsson (to appear), for a more general overview of this construction type in Papuan languages). For the argumentation here, this is irrelevant.

(10) *hatte o yaha wilamuk ulugte embeselug libag*
i. [hat=te o yaha wilat-muk] **u-lug=te**
2SG=TOP house watch stay-2.FUT.IMP **say-SS.SEQ=TOP**
ii. embese-lug
leave-SS.SEQ
iii. la-ibag
go-3.RPST
'"you stay and guard the house" they said, left (him), and went (into the forest)' [N fighting_a_ghost_058]

However, four observations about instances of *ulug* in the corpus are remarkable: First, constructions of reported speech and *ulug* are frequently interpreted and translated in a non-literal way by our Yali research collaborators, i.e., as constructions of extended reported speech. This, we suggest, is remarkable, because the same is *not* true for reported speech followed by a finite speech predicate, as, e.g., in (9). Second, in numerous cases, *ulug* is immediately followed by an inflected form of *uruk* 'to say', i.e., we find a doubling of one and the same speech verb. Third, there are instances where *ulug* occurs in sentence-final position, which, too, is remarkable, given that *ulug* is a medial verb form. And finally, we find instances where *ulug* is used without a speech-related meaning following a noun phrase.

We regard the behaviour of *ulug* in these four contexts as indicative of grammatical change. In the following subsections, we will therefore describe these contexts in more detail. Section 3.1 first takes another look at constructions like (10), which we consider to be the SOURCE CONSTRUCTION in the grammaticisation process of *ulug*, showing that *ulug* undergoes semantic bleaching when it follows extended reported speech. Section 3.2 investigates the doubling of *ulug* and a second speech predicate, providing further evidence for its semantic change. Section 3.3 discusses the use of sentence-final *ulug*, arguing for an interpretation in terms of epistemic stance, and Section 3.4 presents the use of *ulug* as a noun-phrase marker.

3.1 Extended reported speech and *ulug*

As discussed above, *ulug* can be used as a medial SIC with its literal meaning 'to say', and as such it serves to integrate speech events into clause chains expressing a sequence of events. (10) already demonstrated this use, (11) provides another example.

(11) *ye ye ye atukno we nakol o ulug ambirisi*
i. [ye ye ye atuk=no we n-akol o] **u-lug**
yes yes yes over.there=GIV PTL 1SG.GEN-fear PTL **say-SS.SEQ**
ii. ambu-tisi
go.down-3SG.PST
'"yeah, yeah, yeah, (those) over there, I'm scared", he said, and went down'
[C women_chat_1 317]

Reported speech constructions in Yali do not, however, only serve to report what someone said, but are also the default means to express the speaker's or other people's (assumed) thoughts and intentions and thus the reasoning for their actions. We introduced this kind of usage and, following Spronck and Carsatelli (2021), referred to it as *extended reported speech* in Section 2.2. Extended reported speech denoting thoughts or knowing (Example (12)), purpose (Example (13)), and intention (Example (14)), is a very frequent reading of constructions with *ulug* in our data.

(12) *ulug* with extended reported speech, knowledge reading
ar itnoen kilafehek ulug walug lahasareg…
i. [at itno=en kilap-ehek] **u-lug**
3SG DET=ERG go.down-3SG.IMPST **say-SS.SEQ**
ii. wa-lug la-ehesa-teg
take-SS.SEQ go-3PL.IMPST-SS.SEQ
'they knew that he dropped (his hat), and brought it (to him)'
LIT: '"he dropped (his hat)" they said, and brought it (to him)'
[R pear_story_Marthen_b 057]

(13) *ulug* with extended reported speech, purpose reading
we yahiye tia wamin ulug waharuk ari
i. [we yahiye tia wa-min] **u-lug**
PTL frog MED.LOC take-1SG.IMFUT **say-SS.SEQ**
ii. waha-tuk ari
come-INF MED
'he is coming towards those frogs to catch them' [R frog_story 380]

(14) *ulug* with extended reported speech, intentional reading
hik ha <la-> lamin ulug wal halug pok atuk
i. [hik ha la-min] **u-lug**
throw 3SG.OBJ:DIR go-1SG.IMFUT **say-SS.SEQ**
ii. wal ha-lug
try see-SS.SEQ
iii. pok at-tuk
wait become-INF
'he wants to throw it, he tries, (but) then he waits' [D ECOM_sr_KW 152]

Semantically, the translations in Examples (12) to (14) could be regarded as figurative interpretations, deriving more or less transparently from literal readings of an (inner) speech event and a following action ('"I will catch them" he said (to himself), and came' in (13), etc.). As such, we suggest, they can be described as semantic-pragmatic context expansions (cf. Himmelmann 2005: 82) compared to reported speech proper.

Syntactically, on the other hand, a clause chain analysis parallel to examples of reported speech proper ((11) above) is at least possible. Accordingly, we still allocate one line in our one-clause one-line notation for the speech report followed by *ulug* in the exposition of examples in this subsection.

Importantly for our purposes here, the extended interpretations of knowing, purpose and intention above are the original translations of Examples (12) to (14) given by our Yali research collaborators. As such, they can be said to represent a native speakers' metalinguistic knowledge of the constructions, and, remarkably, similar interpretations are notably *absent* for constructions of reported speech without *ulug* in our data. Though we acknowledge that this does not necessarily prove that — given the right context — similar extended interpretations could not be elicited for constructions with a finite speech predicate like, e.g., *imisi* in Example (9), we want to stress the fact that in our corpus extended reported speech does not occur without *ulug*. The reason for this, we suggest, is that prototypically, extended reported speech is a strategy to express the inner motivations for someone's actions, e.g., the protagonists' actions in a narrative. Consider once more Example (12), which stems from a re-telling of Chafe's (1980) pear movie and illustrates this nicely. Here, the protagonists' action of taking the boy's hat and giving it back to him is motivated by their preceding observation that the hat fell down while the boy was riding off with his bike. This kind of reasoning is rather naturally expressed as a sequence of events, and the default syntactic structure for such sequences in Yali is a clause chain. Hence the medial form of the SIC used.

Note that even where a speaker wants to express meanings of extended reported speech in its own right, and not as an explanation of further actions, the syntactic structure of a clause chain is the most typical choice in our data. In such cases, a dummy final clause with the semantically light verb *suruk* 'to do' apparently has to follow the speech report, see Example (15). That is, for extended reported speech, syntactically, the clause chaining structure is the typical choice, even if the final clause is semantically void. Semantically, the construction here starts losing its sequential organisation.

(15) *ulug* followed by a dummy final clause, intentional reading
wisalik lahuk ulug sirikik
i. [wisal-ik la-uhuk] **u-lug**
open-DIR go-1SG.IMFUT **say-SS.SEQ**
ii. **su-tikik**
do-1SG.PST
'I wanted to go and open (the bark)' [E making_net_bag 004]

3.2 Reported speech + *ulug* followed by a second speech verb/SIC

Unlike extended reported speech, reported speech proper can be, and often is, introduced by a simple, inflected form of the verb *uruk* 'to say', as illustrated in Examples (2) and (4) in Section 2.2. Just as frequent, however, is a doubling, where the medial form *ulug* appears in combination with an inflected, final form of *uruk* 'to say', as demonstrated in (16).

(16) Reported speech with *ulug* and final form of *uruk*
at Agusente an nuhe ulug irisimu…
at Agus=en=te [an n-he] **u-lug** **u-tisi=mu**
3SG PN=ERG=TOP 1SG 1SG.GEN-woman **say-SS.SEQ** **say-3SG.PST=DS.SEQ**
'because Agus said "this is my wife"…' [C conversation_1_487]

Less frequently, one also finds instances where the medial verb form *ulug* is followed by a semantically more specific speech predicate, as, for example, the light-verb construction *hiyag uruk* 'to inform' in (17).

(17) Reported speech with *ulug* and a more specific speech verb
ande nebe tot tu himin nuwabuk tehen / ulug hiyag ibag
[an=te n-ebe tot tu ha-min n-wabuk tu-ehen]
1SG=TOP 1SG.GEN-body small PRX see-1SG.IMFUT 1SG.GEN-hit:POT do-2SG.PRS
u-lug **hiyag** **ibag**
say-SS.SEQ **inform** **say:3.RPST**
'"I am small, I could be hit", he informed (him)' [N fighting_a_ghost_031–032]

An analysis of *ulug* and the following speech predicate as a clause chain would imply that the two examples in (16) and (17) involve two speech predicates respectively (i.e. *ulug* plus *irisimu* and *ulug* plus *hiyag ibag*). Semantically, this is unlikely, particularly in the case of (16), where the two predicates are identical and would thus be fully redundant. And indeed, this kind of construction is never translated as two events by native speakers of Yali.

A final important observation is that both construction types — *ulug* + SAY and *ulug* + specific speech predicate — are dominantly used to introduce reported speech proper and hardly ever for extended reported speech. We will come back to this point in Section 4.

3.3 Sentence final *ulug* as a marker of epistemic stance

In a third usage, *ulug* can occur sentence-finally, e.g., turn-finally in a conversation, as in (18), or simply following subjective statements, as in (19).[7]

(18) sentence final *ulug*
Speaker A: *ai we pikit loleg lit ulug*
ai we pikit lol-eg lit **u-lug**
EXCL PTL direct open-ST.PTCP SS.SEQ **say-SS.SEQ**
'oh, **(I think)** it should be opened directly'
Speaker B: *ah pikit loleg lit fug*
ah pikit lol-eg lit fug
EXCL direct open-ST.PTCP SS.SEQ NEG
'ah, it should not be opened directly'
[C conversation_2 0937–0938]

(19) sentence final *ulug*
inibam fanowon / wereg fanoma werehon ibma ulug
i. in-iam fano=on wereg
3PL.GEN-house good=AM EXIST
ii. fano=mu wereg=on ibma **u-lug**
good=LOC EXIST=AM this **say-SS.SEQ**
'their houses are great, **(I think)** they live in a good place'
[C conversation_1 424–425]

These usages of *ulug* can probably best be understood as marking stance, i.e. as a means for speakers to "encode opinions and assessments" (Gray & Biber 2012: 15). More specifically, the usage seems to express (or pragmatically entail) the speaker's uncertainty about the truth value of the preceding proposition, i.e. epistemic stance. It is therefore also not surprising that it particularly often occurs in clauses with potential mood (cf. Riesberg to appear c, Section 3.3).

Semantically, the usage is clearly related to the use of *ulug* in extended reported speech discussed in Section 3.1, i.e. to the expression of thoughts or knowledge in terms of (inner) speech. Consider, for example, the data in (20) and (21). The former is an excerpt from a retelling of the picture book *Frog, where are you?* (Mayer 1969). Here, the speaker describes the scene where the dog puts its head into the jar to look for the frog, because it knows that it has been sitting in there the night before. Functionally, (20) is thus very similar to Example (12), i.e.

7. For transparency we keep the morphological analysis and the glossing consistent as *u-lug* 'say-SS.SEQ', even though here and in other contexts discussed below *ulug* arguably has lost its medial-verb status.

it is motivating a protagonist's actions. The difference is that in (20) no final clause follows *ulug*, which would describe the resulting action.

(20) sentence-final *ulug*
tia horiyeg lahaon ulug
tia horiye-eg laha=on **u-lug**
MED.LOC sit-ST.PTCP do:3SG.IMPST=AM **say-SS.SEQ**
'**(he knows that)** it has been sitting there' [R frog_story 071]

(21) provides a further example in which the speaker describes the intentions of a protagonist (again the dog in *Frog, where are you?*). In contrast to the examples in Section 3.1, however, there is an added stance component of uncertainty on the speaker's part in our research collaborators' interpretation of (21). Note that the clause preceding *ulug* here is also no longer a direct speech report, because, unlike the examples described in Section 3.1, it does not take an internal perspective.

(21) sentence-final *ulug*
kin lendi ambik watmisi ano ulug
kin lendi ambu-ik waha-misi ano **u-lug**
bee above go.down-DIR come-3SG.IMFUT Q **say-SS.SEQ**
'**maybe** he wants to get down the bees up there' [R frog_story 196]

The sentence-final position of *ulug* is syntactically remarkable, because medial verb forms are normally banned from this position (but see Sarvasy 2015, and Gayler 2025, Chapter 3 for possible exceptions). The occurrences in (18) to (21) could thus be analysed as cases of insubordination, i.e., as "the conventionalised main-clause use of what, on prima facie grounds, appear to be formally subordinate clauses" (Evans 2007: 367). Pragmatic entailments of epistemic, evidential or mirative force are not unexpected with insubordination (see, e.g., Evans and Watanabe (2016: 12ff) for such entailments in the Australian language Kayardild). Conventionalising, they could be seen as a further semantic-pragmatic context expansion in addition to what was described in Section 3.1. Similar entailments are neither attested in our data for final SICs (since final clauses are independent main clauses anyway and thus cannot be insubordinated), nor for *ulug* followed by a final clause (i.e. the expected use of a medial clause).

3.4 *Ulug* with noun (and prepositional) phrases

Finally, *ulug* can occur following a noun phrase. Where this is the case, three different interpretations are possible. First, *ulug* can evoke a purposive/benefactive meaning 'for', as illustrated in (22). In this example, the speaker asks her husband to pass her the meal's leftovers, because she wants to feed them to the pigs. It is

conceivable that this use is a further semantic expansion of extended reported speech — a possible more literal translation could thus be 'father, thinking "the pig" I am preparing this'. Interestingly, in these contexts, we sometimes also find an additional use of the postpositions *fahet* or *feyeg* 'for', as in (23). We consider those uses to constitute the bridging construction and take them as evidence that *ulug* has not (yet?) acquired a full postpositional status, but rather still retains traces of speaker subjectivity. A similar purposive use of a quotative marker has been described for Marind in Olsson (2021: 99), though note that the Marind quotative probably has its origin in a manner deictic and not in a speech verb/SIC (Olsson 2021: 97).

(22) *ulug* following an NP
hundu an wam itno ulug mel suruk lahi
hundu an **wam itno u-lug** mel su-tuk lat-ihi
father 1SG **pig DET say-SS.SEQ** PH do-INF do-1SG.IMPST
'father, I am preparing this for the pigs' [C conversation_2_0055]

(23) *ulug* following a PP with *fahet* 'for'
uang ita fahet ulug hubet lahik waharuk ari hia
uang ita fahet u-lug hubet laha-ik waha-tuk ari hia
money DIST for say-SS.SEQ early.morning go.up-DIR come-INF MED that
'he goes up very early for the money' [C conversation_2 0214]

Second, and probably related to the uses in (22) and (23), an NP followed by *ulug* can occur in apposition to another noun phrase. In those cases, rather than denoting the purpose of an action as in (22), the *ulug*-marked NP then specifies the role or function of the second NP, similar to English 'as'. Thus, in (24) the phrase *onggo ulugte*, best translated as 'as a payment', specifies the function of the NP *e anggen* (*tigaen*) '(three) fruits'. Again, an interpretation in terms of extended reported speech, i.e. as 'thinking "payment" he gave them three fruits' is not inconceivable.

(24) *ulug* following an NP; appositional function
onggo ulugte e anggen og / e anggen tigaen og / isehek itano
onggo u-lug=te e anggen og e anggen tiga=en og
charge say-SS.SEQ=TOP tree fruit give tree fruit three=ERG give
isa-ehek ita=no
3PL.OBJ-3SG.IMPST DIST=GIV
'as a payment, he gave them three fruits' [R pear_story_Saulus_Lidya 132–134]

And finally, as illustrated in (25), an NP followed by *ulug* can receive a similative interpretation. Note that in this example, the *ulug*-marked NP 'like uninitiated men' occurs as an afterthought in sentence-final position, a position that is also

available for the other two functions described above, but otherwise not possible for medial verbs, as discussed in Sections 2.1 and 3.3.

(25) *ulug* following an NP
ap inanggom elehon / <wa-> wit ari kilap inaptuk lapag fug / we selama-lama weregma ohore / […] */ ap we kuilon ulug*
ap in-anggom eleg=on wit ari kilap inap-tuk lat-pag fug
man 3PL.GEN-pig NEG=AM ceremony MED enter 3PL.THEME-INF do-3.RPST NEG
we selama-lama wereg=mu oho=te […] **ap we kuilon**
PTL for.good EXIST=DS.SEQ also=TOP **man PTL non.initiated.man**
u-lug
say-SS.SEQ
'those who don't have any pigs cannot enter (their children) for the initiation ceremony. they live (like this) for ever [until they have the same age as Laurens and me], like uninitiated men' [E 1st_initiation 482-487]

We suggest that both, the purposive 'for' and the functional 'as' interpretations of (NP+) *ulug* are grammatical extensions of the intentional uses of extended reported speech discussed in Section 3.1, since, just like extended reported speech, the NP+*ulug* complex can be seen to functionally motivate the following action for a listener. Note, however, that in addition to the semantic and syntactic expansions that were also observed in Sections 3.1–3.3, we are here, furthermore, dealing with a host-class expansion: the host of *ulug* is no longer a (reported speech) clause, but a noun phrase.

The close relationship of SICs and quotative indices more generally, on the one hand, and similatives, on the other hand, is typologically widely attested (cf. among others, Güldemann 2008: 417–422 on various African languages, Guz 2024 on Polish, or Kapitonov and Gentens, this volume, on the Australian language Kunbarlang).[8] Likewise, the development of speech verbs and quotatives into nominal markers is well attested, especially in the African literature. Güldemann lists a number of languages in which "the shape of the quotative gram is close to or identical with a […] versatile nominal marker" (2008: 512). Particularly similar to the Yali case seem to be Dongola, for which Armbruster (1960, as quoted in Güldemann 2008: 512) describes a postposition meaning 'concerning, as to the effect of, for' that grammaticised from a quotative verb, and Baka, where a former speech verb has taken on a number of different grammatical functions, among them one as a nominal marker signalling benefactives and purpose (Kilian-Hatz 1995: 269, 272).

8. But note that the directionality of grammaticisation can be reverse, i.e. from similative to quotative, as, e.g., nicely demonstrated by Kapitonov and Gentens (this volume).

4. Discussion

In the previous section, we described medial and final SICs with the speech verb *uruk* 'to say' in Yali. Both can serve to introduce direct quotes reporting someone's speech — in the case of a medial clause within a sequence of events, in the case of a final clause as the last event of a sequence or as a non-chained event. However, we observed a functional difference between the two types of SICs: the medial SIC *ulug* is frequently used to introduce constructions of extended reported speech as well, while final SICs with *uruk* are attested with reported speech proper only.

Furthermore, we described three usages of *ulug* that are unexpected for medial verbs in Yali: Firstly, *ulug* is frequently followed by a semantically fully redundant final clause using a final form of the same speech verb *uruk*. This usage is semantically unexpected for a chaining construction, which normally denotes two (or more) subsequent events. In this specific case, however, the two verbs do not express a sequence of two events, but refer both to the same speech event instead. Secondly, *ulug* can occur sentence-finally, which is unexpected from a grammatical point of view, since medial clauses canonically should be followed by final clauses. And thirdly, *ulug* can occur as a marker of NPs (or PPs) in Yali, which is not possible for other medial verb forms. Together with the functional difference between medial and final SICs, we argued, these three usages of *ulug* indicate grammatical change. In this section, we will discuss how this change can be captured as a process of grammaticisation in the sense of Himmelmann (2005) (Section 4.1). Following this, we will reflect on the grammatical status of *ulug* in Yali from a synchronic perspective (Section 4.2). Finally, we will shortly discuss the relevance and (in)conclusiveness of synchronic data for the discussion of grammaticisation (Section 4.3).

4.1 The grammaticisation of *ulug* in Yali: A possible cline

Assuming a process of grammaticisation in the sense of Himmelmann (2005), we can describe *ulug* as the grammaticising element, changing in its combination with direct quotes, which are the host class of the process. Both together, *ulug* and the quote, are thus the grammaticising construction. Evidence for their tight association comes from their prosodic realization: prosodic boundaries between *ulug* and the quote preceding it are extremely rare in our data (though not impossible, cf. (17) above and (27) below), i.e., in the vast majority of all cases, the two are realised within the same IU (cf. all other examples with *ulug* in this paper). Note that boundaries following *ulug*, on the other hand, are more common. An example of this is (26) (intentional reading).

(26) Prosodic boundary after *ulug*
Jayapura ambul ***ulugte*** */ lumu tiket onggo paltul ulug ahunente …*
i. Jayapura ambu-ul u-lug=te
PN go.down-1PL.IMFUT say-SS.SEQ=TOP
ii. lu=mu tiket onggo pal-ul u-lug ahun=en=te
there=LOC ticket cost cut-1PL.IMFUT say-SS.SEQ husband=ERG=TOP
'they wanted to go down to Jayapura and then they wanted to buy tickets, and her husband (…)' [C conversation_1_512]

The change of the construction can be described as a semantic-pragmatic context expansion, where next to literal speech events, (assumed) reasoning behind someone's actions, i.e. extended reported speech, can now be expressed as well. This happens within a specific constructional context, namely within clause chains in which a reporting speech event forms a sequence with (an) other event(s). Schematically, the grammaticising construction can thus be represented as in Figure 2.

[[direct quote – *ulug*]$_{\text{grammaticising constr}}$ – (n medial clauses) – final clause]$_{\text{larger constr context}}$

dependency relation

Figure 2. Clause chain as grammaticising construction (source)

In addition to the semantic-pragmatic context expansion, a host class expansion and a syntactic expansion can be observed as well. That is, in the immediate contexts (i.e. construction-internal) consisting of the grammaticising element and its host, more different types of hosts are becoming possible: in addition to direct quotes, *ulug* starts combining with non-quoted passages (stance-marking use, see Section 3.3), as well as with NPs or PPs (Section 3.4). In the larger syntactic context (i.e. construction-external, cf. FIG 2), uses without a following final clause are emerging (again stance-marking use and NP/PP marking, Sections 3.3 and 3.4). All three dimensions of context expansion identified in Himmelmann (2005: 81f) for processes of grammaticisation are thus represented.

From this, a plausible path of grammaticisation can be schematised in Figure 3: the source construction is a clause chain with a literal medial SIC introducing a direct quote as reported speech proper. Examples of this construction were given in (10) and (11), Section 3. A first change is then the semantic expansion of the construction, now allowing uses of extended reported speech as well, as demonstrated in (13) to (15) above. Note that syntactically, this usage of the construction clearly preserves features of a clause chain, evident especially in the apparent tendency to add a final 'dummy' clause, even if there is no following event to be expressed (see Example (15)). In a further expansion, the construction then starts to lose these

syntactic requirements as well, making sentence-final and NP-marking uses possible (see examples in Section 3.3 and 3.4).

stage 1, source: prototypical sequential chain, medial SIC + reported speech proper
stage 2: extended reported speech, syntactically still a chain, but not necessarily sequential
stage 3: two redundant SICs? / sentence-final usage (epistemic) / NP/PP marking

Figure 3. Possible grammaticisation path for Yali *ulug*

In line with Himmelmann's (2005: 81f) argumentation, the semantic-pragmatic expansion takes a key role in the process outlined here: extended reported speech (stage 2), we suggest, is the 'key' expansion that the other uses can be traced back to. This has been argued on functional-semantic grounds at least for both the sentence-final (epistemic) and the NP-marking use in Sections 3.3 and 3.4 (for the latter possibly via the sentence-final use). For the combination of *ulug* and a second, final form of a speech verb, as described in Section 3.2, a similar relation to the use of extended reported speech is somewhat less clear, however (hence the question mark on its placement at stage 3 in Figure 3).

Still, one possible explanation for this use that is at least compatible with our data could be that with its very frequent usage in constructions of extended reported speech, *ulug* is starting to lose its ability to introduce reported speech proper, or at least is becoming ambiguous between the two, such that where the intended meaning is a speech report proper, speakers start feeling the need for adding another speech verb to make this intended meaning clear. In our data, usages that tentatively might fit such a description can indeed be found. An example is (27), where the second speech verb, reinforced by the adverbial expression *ari roho* 'like this', could have a clarifying and/or emphatic function ('he *said* this').

(27) Two semantically redundant SICs: emphasis
nambusul ni uruk lahepma / ulugte ari roho uruk latfag
i. [nambusul ni u-tuk lat-ehep=mu] **u-lug=te**
deceit ? say-INF do-2PL.IMPST=DS.SEQ **say-SS.SEQ=TOP**
ii. ari tu-oho u-tuk lat-fag
MED do-ADV say-INF do-3.RPST
'this is just a fairy tale (LIT: you are telling lies), he was speaking like this'
[N fighting_a_ghost_014]

This explanation would be in line with the fact that also in the absence of additional reinforcing elements like *ari roho*, the very frequent combinations of medial and final forms of *uruk* 'to say' (Example (16) in Section 3.2) almost always express speech reports proper in our data. It would also account for combinations

of *ulug* with a more specific verb of saying, as, e.g., 'to inform' in Example (17) (Section 3.2).

Regardless of whether this proves to be correct, the semantic expansion to the domain of extended reported speech clearly plays an important role in the grammaticisation of the medial SIC *ulug* in Yali. As mentioned in the Introduction, it is well possible that the same is true for other Papuan languages as well: extended reported speech requiring a specific medial form — rather than a final form — as SIC has been described elsewhere in the literature before. However, to the best of our knowledge, with the exception of the pioneering work by Reesink (1993), this has never been discussed explicitly in the Papuan literature. More thorough typological investigations are thus needed to gain a better idea of whether extended reported speech as the driving force in the grammaticisation of speech verbs is a more widespread phenomenon cross-linguistically.

Finally, in closing this subsection, it should be stressed here that, as we are dealing with synchronic data in this paper, the traced 'path' in FIG 3 should not be interpreted as a literal step by step progression. Rather, it can be viewed as a functional-semantic or logical tracing of the relation between different usages, which are, however, all simultaneously active in present-day Yali. This, we suggest, is very well compatible with the idea of grammaticisation as a process of expansion advocated in Himmelmann (2005).

4.2 The gram: Is *ulug* a quotative marker?

One of the cross-linguistically most expected outcomes in the grammaticisation of SICs in combination with direct quotes is a dedicated quotative marker, i.e., a grammatical item indicating reported speech (cf., inter alia, Güldemann 2008; Klamer 2000). As discussed in Section 1.3, such a grammatical marker would be expected to have lost its clausal status, and consequently also the ability to introduce the reported speaker and the reported addressee (cf. also Klamer (2000) on the 'loss of argument structure' of former speech verbs in Tukang Besi and Buru). Whether this is also the case for Yali *ulug*, and if so, for which stage(s) of the path proposed in Section 4.1, is, however, not fully straightforward to tell from our data.

One reason why it is difficult to determine the clausal status of *ulug* is that also for non-grammaticised medial clauses, the omission of arguments, and particularly of the subject, is the norm rather than an exception.[9] That is, also as an 'ordinary' medial verb, *ulug* would be expected to frequently occur without overt

9. This dispreference for overt arguments in chaining structures is discussed in the Papuan literature under the term *distribution*, e.g. de Vries (2005, 2006), Himmelmann & Riesberg (to appear).

realisation of the speaker referent. A possible cue for an actual ban against overt subject arguments of *ulug* in combination with extended reported speech could be that in the very rare cases where the reported speaker (or, more accurately, the intender) is realised as an overt NP, these NPs tend to not show ergative marking and be separated from the following combination of quote and *ulug* by a prosodic boundary. An example of this was (5), repeated here for convenience as (28). This is in contrast to those cases where the speaker is realized within final SICs, which tend to be ergative marked and may occur within the same IU as the quote (and the speech verb). One could thus argue that the clause-initial NP in (28) is syntactically less integrated into the main clause than realisations of the reported speaker with final SICs.

(28) Extended reported speech, intentional reading
mene itno / e fene lahamin ulug / pululu hohobo suruk ari
mene itno [e fene laha-min] u-lug pululu hohobo su-tuk ari
dog DET tree on go.up-1SG.IMFUT say-SS.SEQ sound scratch do-INF MED
'the dog wants to climb the tree and is scratching (the trunk)'
LIT: 'the dog says "I will climb the tree" and is scratching (the trunk)'
[R frog_story_Silpa_Fince 099–101]

Note, however, that the same does not hold for *ulug* followed by a second speech verb (with reported speech proper). An example was given in (16), Section 3.2, where the NP referring to the reported speaker is ergative marked and realised within one IU with the speech report and the two speech verbs. If we assume *ulug* to be a quotative marker, the reported-speaker NPs in these cases would have to be analysed as the subject of the second (final) speech verb.[10]

More compelling evidence for a quotative status, however, can perhaps be found from realisations of other syntactic constituents: both with extended reported speech and in the combination of *ulug* and a second final speech verb it is possible to realise different types of constituents, e.g. oblique arguments (e.g. the addressee) or modifying adverbials before the final verb of the construction, which corresponds to the expected slot for such elements if they are arguments, adjuncts, or modifiers to the second verb. An example with a modifying adverbial was presented as (27), an example with an overt realisation of the reported addressee is given in (29).[11] In contrast to this, it is never the case that a modi-

10. A reviewer suggested that the inability to take negation could be further evidence that *ulug* is no longer a medial SIC. Indeed, there is only one instance in the corpus, where extended reported speech + *ulug* is followed by a negation. Interestingly, this is the main clause negator *fug*, and not *eleg*, which would be expected if *ulug* was a medial verb (cf. Riesberg to appear b, Section 3.2). We take this as further tentative evidence for grammatical change, but as negation in Yali is not yet well understood, we leave this issue for further research.

fier or the addressee argument precede *ulug*. That is, both with extended reported speech and in the construction with two redundant speech verbs, the quote and *ulug* are always directly adjacent. This is particularly telling in the case when another speech verb follows: taken as a literal medial verb, the first *ulug* in (29), for example, should semantically have the same addressee referent as *iyuk*. Likewise, the adverbial modification 'like this' in (27) would apply to *ulug* as much as to the final speech predicate *uruk latfag*. The striking absence of such modification can therefore be taken as evidence for a tight association of *ulug* and the quote, and thus favours a description as a quotative marker for this usage.

(29) Overt realisation of addressee preceding the second speech verb
muluk Ayabiye / lahakek ulug / hat hikni hisinga fam / e / iyuk o ulug ibag
[[muluk Ayabiye laha-ehek] **u-lug** hat h-ikni
k.o.lizard PN go.up-3SG.IMPST **say-SS.SEQ** 2SG 2SG.GEN-father
h-isinga fam e **u-iyuk** o] u-lug ibag
2SG.GEN-mother to INTJ say-2.FUT.IMP PTL say-SS.SEQ say:3.RPST
'"say to your fathers and mothers "muluk Ayabiye went up"" he said'
[N land_slide_ayabiye_040]

Again, given that we are dealing with spontaneous data, both observations, the dispreference for subject arguments of *ulug* in constructions of extended reported speech and the strict adjacency of *ulug* and the quote — even in the presence of semantic modification and arguments of the speech event — cannot give certainty that overt arguments of *ulug* are (becoming) ungrammatical in Yali. Nevertheless, we suggest that, together with the general tendency to avoid overt (subject) arguments in clause chaining structures, they present a usage situation that makes the loss of argument slots for *ulug* very *possible*. In this sense, we are describing favourable usage conditions for the emergence of a quotative marker in Yali.

At the same time, in its sentence-final usage (epistemic, Section 3.3), as well as a marker of NPs/PPs (Section 3.4), *ulug* is already leaving the domain of (direct) quoting: As an epistemic stance marker, *ulug* combines with subjective statements, which, despite being functionally still reminiscent of quotes, do not involve the internal perspective that we find in speech reports. As a marker of NPs/PPs, no quotes at all are associated with *ulug* anymore, even if here, too, it is easy to imagine extended reported speech to be the original literal meaning.

11. Note that Example (29) shows a double quoting structure, where a quote is embedded into another quote, as indicated by the bracketing. The speech event with overt addressee in question *hat hikni hisinga fam* 'to your fathers and mothers' is the embedded one, i.e. the first complex of quote plus *ulug* plus addressee plus final speech verb, here the future imperative *iyuk*.

These two usages, therefore, are best described as emergent further extensions of *ulug* beyond the status of a quotative marker.

4.3 Usage patterns in synchronic language data, or what does it mean to 'take spoken language seriously'?

In the spirit of Himmelmann (1998) and other advocates of the multifunctionality of language documentation corpora (e.g. Woodbury 2011; Schnell 2018; Rice 2018, among many others), we set out to investigate the grammaticisation of the Yali medial SIC *ulug* based on synchronic spoken language data. This approach of relying on a synchronic corpus of spoken speech naturally leaves certain questions unanswered. For one, and as mentioned in Section 3.1, we cannot conclude for sure whether extended reported speech is a metaphorical usage of reported speech constructions, or whether the construction quote+*ulug* has two meanings, extended reported speech and reported speech proper. Vice versa, we also cannot know whether an extended reported speech interpretation with a finite (final) SIC is in principle ruled out, or in elicitation contexts still possible. In the same vein, we cannot say whether epistemic stance is an obligatory meaning component of the sentence-final use of *ulug* depicted in Section 3.3, or rather a frequent pragmatic entailment. And finally, it is also not clear whether, or in which of its usages, *ulug* has lost the grammatical ability to introduce speaker arguments and/or where it retains its syntactic status as a SIC.

Despite these uncertainties, we were able to depict a usage situation that makes the grammaticisation of constructions with medial SICs possible, or even likely: we showed that extended reported speech is a — or rather *the* — typical interpretation for medial, but not for finite SICs in *actual use;* epistemic stance is a typical interpretation of sentence-final, but not chain-internal occurrences of *ulug;* and both prosodic boundaries and overtly realised constituents in between direct quotes and *ulug* (host and gram of our process) are untypical in actual use. The latter favours their becoming "automated as a single processing unit", citing Bybee (2003: 603) again. The former two favour the tighter association of the respective interpretations with one, but not the other construction. In this way, and by 'taking spoken language seriously', we hope to have offered an interesting perspective to grammaticisation theory in presenting an example of *how* semantic-pragmatic context expansion arises.

5. Summary and conclusion

In this paper, we investigated the grammaticisation of the medial SIC *ulug* in Yali. We described the different usages of this form in a corpus of spontaneous speech and discussed where they suggest emerging change from the original function as a medial verb introducing (proper) reported speech events in a sequential chain of events. Given the nature of the constructions that we observed, we argued for a process of grammaticisation in the sense of Himmelmann (2005). Its output, we suggested, can be described as an emerging quotative marker with two further extensions as a marker of epistemic stance and as a (versatile) noun-phrases marker.

As one step in the process, we described the emergence of constructions of extended reported speech, with especially intentional, but also e.g. purposive readings. Since the literature suggests similar processes might be underway in other Papuan languages as well, we therefore hope this paper to be informative to the investigation of Papuan languages more widely.

Funding

We gratefully acknowledge the financial support from the German Research Foundation (DFG) for our research within the project *Bo5 Prominence related structures in Austronesian symmetrical voice and Papuan languages* of the Collaborative Research Centre SFB1252 *Prominence in Language* (Project-ID 281511265) at the University of Cologne. The corpus data used in this chapter has been collected during field work funded by the Volkswagen Foundation (DoBeS, 2012–2016), the Australian Research Council (Centre of Excellence for the Dynamics of Languages, 2017–2020), and the German Research Foundation (SFB1252, 2021–2025).

Acknowledgements

We want to express our sincere gratitude to the Yali speech community who supported our research, and to all the Yali speakers who contributed their stories and language use. Our special thanks goes to our Yali research collaborators, whose expertise and linguistic knowledge have played a foundational role in this paper.

We would also like to thank Uta Reinöhl and two anonymous reviewers for their detailed and helpful comments on an earlier version of this chapter. The chapter notably improved due to their constructive input. Thanks to Matthew Micyk for proofreading the manuscript. All remaining errors are, of course, our own.

We are very grateful for having (had) the opportunity to work with and learn from Nikolaus, who has the capacity of taking not only spoken language, but also collaborators and students seriously. Not all supervisors have the patience and courage to let their students find their own perspectives. His dedication, his critical — but always fair and constructive — assessments, his curiosity and his enthusiasm for venturing into (academically) unknown territory will always be an inspiration to us.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following abbreviations are used:

AM	attributive marker	PH	place holder
CAUSE	cause	PN	proper noun
DIR	directedness	POT	potential
DS	different subject	PRX	proximal
EXCL	exclamation	PTL	(discourse) particle
EXIST	existential	RPST	remote past
GIV	given	SEQ	sequential
IMFUT	immediate future	SS	same subject
IMPST	immediate past	ST	stative
MED	medial	THEME	theme

References

Armbruster, Charles H. 1960. *Dongolese Nubian. A Grammar*. Cambridge: Cambridge University Press.

Berghäll, Liisa. 2015. *A Grammar of Mauwake*. Berlin: Language Science Press.

Boersma, Paul, and David Weenink. 2021. *Praat. Doing Phonetics by Computer*. Version 6.1.50. www.praat.org

Bromley, Myron. 1981. *A grammar of Lower Grand Valley Dani*. Canberra: Australian National University.

Bybee, Joan. 2003. "Mechanisms of Change in Grammaticization: The Role of Frequency." In *The Handbook of Historical Linguistics*, ed. by Brian D. Joseph, and Richard D. Janda, 602–623. Oxford: Blackwell.

Chafe, Wallace. 1994. *Discourse, Consciousness and Time*. Chicago/London: The University of Chicago Press.

Chafe, Wallace. 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex.

Deibler, Ellis. 1971. "Uses of the Verb 'to Say' in Gahuku." *Kivung* 4 (2): 101–110. https://www.langlxmelanesia.com/llm-archive

Deutsche, Guy. 2000. *Syntactic Change in Akkadian. The Evolution of Sentential Complementation*. Oxford: Oxford University Press.

Deutscher, Guy. 2012. "The Grammaticalization of Quotatives." In *The Oxford Handbook of Grammaticalization*, ed. by Bernd Heine, and Heiko Narrog, 646–655. Oxford: Oxford University Press.

de Vries, Lourens. 1990. "Some Remarks on Direct Quotation in Kombai." In *Unity in Diversity*, ed. by Harm Pinkster, and Inge Genee, 291–310. Dordrecht: Foris Publication.

de Vries, Lourens. 1992. "Quotative Constructions and Translation in Papuan Languages." *The Bible Translator* 43 (3): 333–342.

de Vries, Lourens. 2005. "Towards a Typology of Tail–Head Linkage in Papuan Languages." *Studies in Language* 29 (2): 363–384.

de Vries, Lourens. 2006. "Areal Pragmatics of New Guinea: Thematization, Distribution and Recapitulative Linkage in Papuan Narratives." *Journal of Pragmatics*, 38 (6): 811–828.

Evans, Nicholas. 2007. "Insubordination and Its Uses." In *Finiteness: Theoretical and Empirical Foundations*, ed. by Irina Nikolaeva, 366–431. Oxford: Oxford University Press.

Evans, Nicholas. 2013. "Some Problems in the Typology of Quotation: A Canonical Approach." In *Canonical Morphology and Syntax*, ed. by Dunstan Brown, Marina Chumakina, and Greville G. Corbett, 66–98. Oxford: Oxford Academic.

Evans, Nicholas, and Honoré Watanabe. 2016. "The Dynamics of Insubordination." In *Insubordination*, ed. by Nicholas Evans, and Honoré Watanabe, 3–38. Amsterdam/Philadelphia: John Benjamins.

Gayler, Katharina. 2025. Larger Discourse Units in Linguistic Structure of Yali Narratives. Doctoral Dissertation, University of Cologne.

Gray, Bethany, and Douglas Biber. 2012. "Current Conceptions of Stance." In *Stance and Voice in Written Academic Genres*, ed. by Ken Hyland, and Carmen Sancho Guinda, 15–33. London: Palgrave Macmillan.

Güldemann, Tom. 2008. *Quotative Indexes in African Languages: A Synchronic and Diachronic Survey*. Berlin/New York: Mouton de Gruyter.

Guz, Wojciech. 2024. "Quotative Uses of Polish Similative Demonstratives." *Linguistics* 62 (3): 541–575.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. *Glottolog 5.2*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://glottolog.org, Accessed on 2025-10-21.)

Healey, Phillis. 1964. "Teleéfoól Quotative Clauses." *Papers in New Guinea Linguistics* 1: 27–34.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36 (1): 161–195.

Himmelmann, Nikolaus P. 2005. "Gram, Construction, and Class Formation." In *Wortarten und Grammatikalisierung, Perspektiven in System und Erwerb*, ed. by Clemens Knobloch, and Burkhard Schaeder, 79–93. Berlin/New York: De Gruyter.

Himmelmann, Nikolaus P. 2014. "Asymmetries in the Prosodic Phrasing of Function Words: Another Look at the Suffixing Preference." *Language* 90 (4): 927–960.

Himmelmann, Nikolaus P. 2022. "Prosodic Phrasing and the Emergence of Phrase Structure." *Linguistics* 60 (3): 715–743.

Himmelmann, Nikolaus P., and Sonja Riesberg. To Appear. "Discourse Patterns and Emerging Grammar in Papuan Languages." In *The Oxford Guide to the Papuan languages*, ed. by Nicholas Evans, and Sebastian Fedden. Oxford: Oxford University Press.

Himmelmann, Nikolaus P., Meytal Sandler, Jan Strunk, and Volker Unterladstetter. 2018. "On the Universality of Intonational Phrases: A Cross-Linguistic Interrater Study." *Phonology* 35 (2): 207–245.

Kilian-Hatz, Christa. 1995. *Das Baka: Grundzüge einer Grammatik aus der Grammatikalisierungsperspektive*. Cologne: Institut für Afrikanistik, Universität zu Köln.

Klamer, Marian. 2000. "How report verbs become quote markers and complementizers." *Lingua* 110: 69–98.

Mayer, Mercer. 1969. *Frog, Where are You?* New York: Dial Books for Young Readers.

Olsson, Bruno. 2021. *A Grammar of Coastal Marind*. Berlin/Boston: De Gruyter Mouton.

Pawley, Andrew. K. 2005. "The Chequered Career of the Trans New Guinea Hypothesis: Recent Research and its Implications." In *Papuan Pasts: Cultural, Linguistic and Biological Histories of the Papuan-Speaking Peoples*, ed. by Andrew K. Pawley, Robert Attenborough, Jack Golson, and Robin Hide, 67–101. Canberra: Pacific Linguistics.

Pawley, Andrew K. 2009. "Greenberg's Indo-Pacific Hypothesis: An Assessment." In *Discovering History Through Language: Papers in Honour of Malcolm Ross*, ed. by Bethwyn Evans, 153–180. Canberra: Pacific Linguistics.

Reesink, Geer. 1993. ""Inner Speech" in Papuan Languages." *Languages and Linguistics in Melanesia* 24: 217–225.

Reesink, Geer. 2019. "Quotations form a Recursive Discourse." *Linguistic Typology* 23 (1): 229–235.

Reesink, Geer, and Nicholas Evans. To Appear. "Reported Speech and Thought in Papuan Languages." In *The Oxford Guide to the Papuan Languages*, ed. by Nicholas Evans, and Sebastian Fedden. Oxford: Oxford University Press.

Reinöhl, Uta, and Antje Casaretto. 2018. "When Grammaticalization Does NOT Occur: Prosody-Syntax Mismatches in Indo-Aryan." *Diachronica* 35 (2): 238–276.

Rice, Sally. 2018. "Reflections on Documentary Corpora." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 157–172. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24817

Riesberg, Sonja. To Appear a. "Reported Speech in Yali." In *Reported Speech in the World's Languages: A Comparative Handbook*, ed. by Tatiana Nikitina, Stef Spronck, Denys Teptiuk, and Anna Bugaeva. Berlin/Boston: De Gruyter Mouton.

Riesberg, Sonja. To Appear b. "Verbal Complex Predicates in Yali." In *Verbal Complex Predicates: A Comparative Handbook*, ed. by Walter Bisang, Alexander Francois, Andrej Malchukov, and Martine Vanhove. Berlin/Boston: De Gruyter Mouton.

Riesberg, Sonja. To Appear c. "Potential Mood and Apprehensional Modality in Yali." In *Linguistic Worlds and Social Minds: What Language Documentation and Linguistic Diversity Can Tell Us About the Nature and Origin of Language*, ed. by Wayan Arka, Danielle Barth, and Henrik Bergqvist.

Riesberg, Sonja, Kristian Walianggen, and Siegfried Zöllner. 2012–2016. *DoBeS Documentation Summits in the Central Mountains of Papua*. The Language Archive MPI Nijmegen, http://dobes.mpi.nl

Riesberg, Sonja, and Bruno Olsson. To Appear. "Coverb Constructions in Papuan Languages." In *The Oxford Guide to the Papuan Languages*, ed. by Nicholas Evans, and Sebastian Fedden. Oxford: Oxford University Press.

Sarvasy, Hannah. 2015. "Breaking the Clause Chains; Non-Canonical Medial Clauses in Nungon." *Studies in Language* 39 (3): 664–696.

Sarvasy, Hannah. 2017. *A Grammar of Nungon, a Papuan Language of Northeast New Guinea*. Leiden/Boston: Brill.

Saxena, Anju. 1988. "On Syntactic Convergence: The Case of the Verb 'Say' in Tibeto-Burman." *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics society*. Linguistic Society of America: published online. 375–388.

Schnell, Stefan. 2018. "Reflections on the Role of Language Documentations in Linguistic Research." In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 173–182. Honolulu: University of Hawai'I Press. http://hdl.handle.net/10125/24818

Spronck, Stef, and Daniela Casartelli. 2021. "In a Manner of Speaking: How Reported Speech May Have Shaped Grammar." *Frontiers in Communication* 6: 624486.

Woodbury, Anthony C. 2011. "Language Documentation." In *The Cambridge Handbook of Endangered Languages*, ed. by Peter K. Austin, and Julia Sallabank, 159–186. Cambridge: Cambridge University Press.

CHAPTER 30

Of Malays, Moluccans and Papuans

The development and spread of Malay in Papua

Antoinette Schapper
Lacito-CNRS | The Australian National University

Little is certain about the origins of Papuan Malay. The literature offers two opposing viewpoints: one is that Papuan Malay arose out of centuries of Malay use in New Guinea connected with indigenous trade networks centred on Ternate and the North Moluccas, the other is that Papuan Malay arose very late as part of colonial policies in the 20th century. The chapter presents the first examination of these conflicting claims on the basis of an in-depth analysis of the lexicon of Papuan Malay and Malay loanwords along the north coast of New Guinea. Papuan Malay is shown to have a significant lexical component from Ternate, like other vehicular Malays in the region. Despite this, the evidence as a whole supports the theory of a recent, not historically distant, origin for Papuan Malay.

Keywords: Papuan Malay, Ternate, lexical documentation, historical linguistics, loanwords

1. Introduction

Today, Papuan Malay is one of the best described Malay varieties in eastern Indonesia. The last twenty years has seen significant effort go into the documentation and description of Papuan Malay as it is spoken in northwest New Guinea (e.g., Saragih 2012; Kluge 2014, later published as Kluge 2017; Himmelmann & Riesberg 2012–2016). Despite this extensive documentation, little is certain about the origins of Papuan Malay.

The literature contains significant disagreements about the timing and mode of introduction of Malay to the island of New Guinea. On the one hand, scholars like Rowley (1972), Roosman (1982) and Paauw (2009) take the view — one I have often heard repeated — that Malay was in use for centuries along the coasts of Papua as part of regional trade networks:

https://doi.org/10.1075/slcs.240.30sch
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

> Malay has had a presence in the coastal areas for a long time. Malay-speaking traders visited what is now Indonesian Papua (and beyond) before the European colonial era. The Raja Ampat islands to the northwest of Papua and the north coast of the Bird's Head region in northwestern Papua were included in the lands under the influence of the Sultan of Tidore before the first Europeans visited. Malay traders were known in other coastal areas, particularly in the north and west, and even in the eastern half of the island. (Paauw 2009: 71)

> Papuan Malay initially developed through contacts (in coastal areas only) with the sultanates of Tidore and Ternate in the 17th and 18th centuries, followed by trade contacts with the north Moluccans and Ambonese in the 19th century, as well as religious propagation (from Ambon and the central Moluccas) and Dutch colonial administration in the 19th and 20th centuries, and finally, Indonesian government administration, education and mass media after 1963. There appear to be regional variations due to the differing forms of the contact language which brought Malay to Papua, with the area around Cendrawasih Bay speaking a variety closer to North Moluccan Malay, and the southern Bird's Head and Bomberai peninsula speaking a variety closer to Ambon Malay. (Paauw 2009: 307)

On the other hand, Seiler (1985) examines the historical record and comes to the conclusion that almost no Malay was used along the north coast of New Guinea before the 20th century. Additionally, he shows that the Malay element in Tok Pisin is not due to trading contacts along the coasts, but originates in use of Malay by labourers in the plantations of the old German protectorate in northeast New Guinea. Kluge (2017: 42–47) offers a forceful restatement of Seiler's (1985) findings, providing more historical documentation for the lack of Malay in New Guinea throughout much of the 19th century and concluding that Malay only developed a role in Papua in the 20th century.

This chapter seeks to clarify historical aspects of Malay's role as a contact language in Papua and the emergence of Papuan Malay. To date, the focus of the discussion on the history of eastern Malays has been on aspects of their grammar and highly emblematic grammatical items such as pronouns and negators (see, e.g., Collins 1980; Paauw 2009). Additionally, research has shown that as Malay became the preeminent vehicular language in Papua, local languages have had transformative effects on the language's morphosyntax (Gil 2024). Because my concern is with the influences that contributed to the initial formation of Papuan Malay, I limit myself in this chapter to looking at lexical data. Although observations are scattered throughout the literature, there has not been a systematic treatment of the distinctive lexical characteristics of Papuan Malay and the appearance of Malay borrowings in Papua with the view to analysing what these tell us about the history of Malay usage in the region. In this chapter, I restrict discussion to the role and manifestations of Malay in the northern part of the Papuan area, extend-

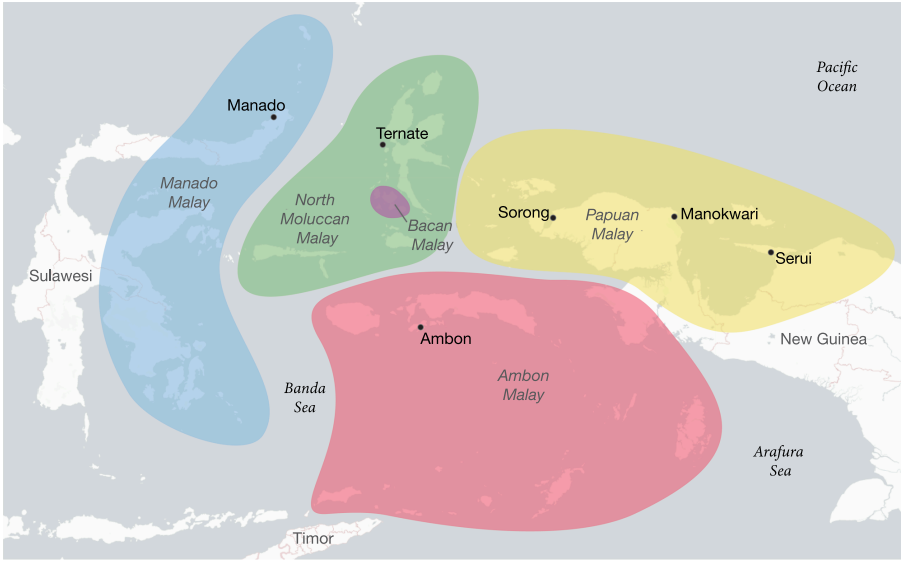
ing from Sorong in the west, over the Bird's Head and through Cenderawasih Bay, along the north coast of New Guinea to the Humboldt Bay area. It is from this region that the current documentation of Papuan Malay originates, and chronological historical documents provide insights into the timing of the introduction of Malay. Other varieties of Malay along the south coast of New Guinea, including the southern Bird's Head and the Bomberai peninsula, look to have a range of different lexical influences which warrant separate historical investigation.

The chapter is structured as follows. Section 2 shows that much of the characteristic vocabulary in Papuan Malay originates in Ternate, thus placing it within the network of Malays in eastern Indonesia that were heavily influenced by the North Moluccas. Section 3 argues that there is little to support the idea that Papuan Malay emerged as a result of a long history of trade contact with the Moluccas. Section 4 concludes.

2. Relationship of Papuan Malay to other Malays

Map 1 sets out the broad geographical areas over which the named Malay varieties treated in this chapter are dispersed. The map includes the locations of key towns named in this chapter. These are often the places in which the different Malay varieties first arose; the varieties are frequently named for these centres. It is important to bear in mind that the representation on Map 1 is a radically simplified picture of the linguistic reality. Dividing neighbouring islands into different Malay blocks is an idealisation which does not do justice to the complexities of Malay usage and how varieties shade into one another. Within a single block, there is also much variation to be found. For instance, although the Malay spoken in Aru is not the same as the Malay spoken in Ambon, highly salient grammatical words of Aru Malay originate in Ambon Malay, e.g., *seŋ* 'no, not' and *lai* 'again'. Such features are used to determine the blocks in Map 1.

This chapter is concerned with understanding the lexical relationships between Papuan Malay and the other varieties of Malay marked on Map 1. As used here, a "lexical relationship" is simply where two or more Malay varieties have shared lexemes of a distinctive character that are not found elsewhere. Where shared lexemes that can be shown to have come from a particular localised source are found, I take this to indicate that one variety of Malay has been lexically influential in the formation of another variety. Whilst this is a crude metric (see Adelaar & Prentice 1996 on the difficulties of classifying Malay varieties), it is fairly uncontroversial to say that looking at lexemes does provide some insight into the historical influences that were present at the time Malay varieties were being formed. However, it does not tell us about how different Malay varieties are from one another.



Map 1. Malay varieties discussed in this chapter

There is strikingly little substance on the lexical influences of Papuan Malay in the existing literature. Donohue (2003, n.p.) writes: "The northern varieties [of Papuan Malay] show a clear influence from M[a]nado Malay / North Moluccan Malay [...], in lexical items such as *kelemarin* 'yesterday'". Without any addition of data, Paauw (2009:73–75) repeats Donohue's (2003) statement and adds that influence from Ambon Malay can also be perceived in Papuan Malay, but concedes: "The amount of influence on the northern variety or varieties from North Moluccan Malay and Ambon Malay is still unclear". Kluge (2017:11, 46) reiterates the observations of both Donohue and Paauw, but seems to express scepticism about Donohue's claim, noting that the form he cites, *kelemarin*, was not present in her corpus of materials. In this section, I aim to provide as complete as possible a review of the lexical influences on the Papuan Malay lexicon.

From my own investigations, it is apparent that the defining lexical relationship that links Papuan Malay with the other Malays of the region is the presence of lexemes that ultimately go back to Ternate, the Papuan language spoken on the island of the same name. Although the importance of Ternate is widely recognised, there has to date been no research into understanding the extent and nature of Ternate lexical influence on the different Malay varieties (Paauw 2009:73). Table 1 sets out more than sixty Ternate lexemes that I have identified in Papuan Malay. These are often also attested in North Moluccan Malay (also known as Ternate Malay), or another regional Malay variety, such as Bacan Malay, Manado Malay and/or Ambon Malay.

Table 1. Papuan Malay lexemes originating in Ternate

Ternate	Papuan Malay	Other eastern Malay varieties
belo 'pole, mooring pole'	*belo* 'punting pole'	North Moluccan Malay *belo* 'stake, pole', Ambon Malay *belo* 'stakes of a fish trap in the sea'
bete 'taro, *Colocasia esculenta*'	*bete* 'taro'	North Moluccan Malay, Manado Malay *bete* 'taro'
bia 'shellfish (generic term)'	*bia* 'shell, shellfish'	North Moluccan Malay, Ambon Malay, Manado Malay *bia* 'shell, mollusc'
bibi 'belly out, sail full of wind; pufferfish'	*porobibi* 'pufferfish'	Ambon Malay *purut bibi* 'potbelly', *poro bibi* 'pot bellied', Manado Malay *purut bibi* 'potbelly'
bilolo 'hermit crab'	*bilolo* 'hermit crab'	Manado Malay *bilolo* 'sea snail'
bobara 'trevally, *Carangidea*'	*bobara* 'trevally, fish sp'	North Moluccan Malay *bobara* 'fish sp', Ambon Malay *bubara* 'smaller jack fish'
bobatu 'kind of plant the roots of which are used to poison fish, *Millettia sericea*'	*bobatu* 'fish poison'	
bobo 'nipa palm, *Nypa fruticans*'	*bobo* 'nipa palm wine'	North Moluccan Malay *bobo*, Manado Malay *bobo* 'nipa palm, *Nypa fruticans*'
bori 'plant the leaves of which contain a drug which stupefies fish, *Anamirta cocculus*'	*bore ~ bori* 'derris root'	Bacan Malay *bori* 'fish poison (from a root)', Ambon Malay *bore* 'vine (the root of which is used as a medicine and fish poison), *Derris elliptica*', Manado Malay *akar bori* 'root of *Millettia sericea*, thrown into the sea to stun the fish'
cafarune 'untidy, filthy, flith'	*caparuni* 'messy'	Ambon Malay *caparuni ~ caparune* 'filthy', Manado Malay *caparuni* 'dirty; obscene, immoral'*
cidako ~ cidaku 'loincloth'	*cidako ~ cidaku* 'loincloth'	Ambon Malay *cidaku* 'loincloth', Manado Malay *cidako* 'loincloth, diaper'
cuki 'genitals', *maku-cuki* 'fornicate'	*cuki* 'fuck'	North Moluccan Malay *cuki mai* 'damned', Ambon Malay *cuki* 'fuck', *baku-cuki* 'have sexual intercourse, copulate', *cuki mai* 'vulgar exclamation of frustration and disgust'**
dabu-dabu 'side-dish of chillies, onions, tomatoes and salt'	*dabu-dabu* 'side dish of chillies'	North Moluccan Malay, Manado Malay *dabu-dabu* 'side dish of raw chillies, tomatoes, and salt in lemon juice'

Table 1. *(continued)*

Ternate	Papuan Malay	Other eastern Malay varieties
dalise 'k.o. fish, *Lutjanus* sp.'	*dalisi* 'k.o. red fish, *Lutjanus* sp.'	Bacan Malay *dalise* 'k.o. fish, snapper, red emperor, *Lutjanus sebae*'
deho 'mackerel tuna'	*deho* 'k.o. fish, tuna'	
dodeso 'snare, trap, noose'	*dodeso* 'snare'	North Moluccan Malay *dodeso* 'snare for catching birds', Ambon Malay *dodeso* ~ *dudesu* 'any snare employing a noose and trigger', Manado Malay *dodeso* 'snare, trap'
jubi 'bow, shoot an arrow', *jubi-jubi* 'bow, speargun'	*jubi* 'bow and arrow'	Manado Malay *jubi-jubi* 'bow for an arrow'
gaba 'dried midrib of palm frond, sago leaf stems which, after drying, are used as (house walls, ceilings, etc)'	*gaba-gaba* 'midrib of palm frond'	North Moluccan Malay *gaba-gaba* 'sago stem, (wall made of) midribs of sago palm fronds', Ambon Malay *gaba-gaba* 'pithy wood of a dry sago branch'
gaba-gaba 'lungs'	*hati gaba* 'lungs'	Ambon Malay, Manado Malay *hati gaba* 'lungs'
gaca 'k.o. fish, *Lutjanus* sp.'	*gaca* '*Lutjanus* sp.'	Bacan Malay *gaca* 'k.o. fish, hussar or sea-perch, *Lutjanus amabilis*', Ambon Malay *gaca* 'k.o. fish, *Lutjanus* sp.'
galala 'k.o. tree, *Erythrina* spp'	*galala* '*Erythrina* sp'	Ambon Malay *galala* '*Erythrina picta*', Manado Malay *galala* 'k.o. tiger claw'
galema 'k.o. ray, manta ray'	*galema* ~ *galemaŋ* 'k.o. large ray'	
ganemo ~ *ganemu* ~ *ganem* '*Gnetum gnemon*'	*gnemo* ~ *ganemon* '*Gnetum gnemon*'	North Moluccan Malay *genem*, Ambon Malay, Manado Malay *ganemu* '*Gnemon gnetum*'
gara 'tease, pester'	*gara* 'irritate'	North Moluccan Malay *gara* 'feign', Manado Malay *gara* 'joke, bother'***
gau 'piece of the young leaf of the areng palm used as cigarette paper'	*gau* 'dried leaves used as cigarette paper'	Bacan Malay *gau* 'paper cigarette wrapper', Manado Malay *gau* 'dried and scraped leaf of the aren palm in which tobacco is rolled into a cigarette'

Table 1. *(continued)*

Ternate	Papuan Malay	Other eastern Malay varieties
gete-gete 'k.o. freshwater fish'	*gete-gete* 'k.o. freshwater fish'	Ambon Malay, Manado Malay *gete-gete* 'k.o. freshwater fish'
gofasa 'k.o. tree, *Vitex* spp'	*gopasa ~ gupasa* 'k.o. tree, *Vitex* spp'	North Moluccan Malay *gofasa*, Ambon Malay *gofasa ~ gopasa* 'k.o. tree, *Vitex* spp'
gomo '*Artocarpus incisa*'	*gomo ~ gomu* 'k.o. breadfruit'	Bacan Malay *gomo* 'k.o. breadnut, *Artocarpus altilis*', Ambon Malay *gomu* '*Artocarpus incisa*'
goraŋo ~ guraŋo 'shark'	*goraŋo ~ guraŋo* 'shark'	North Moluccan Malay *guraŋo*, Manado Malay *goraŋo* 'shark'
gosi 'egg, testicle'	*gosi ~ gusi* 'penis'	North Moluccan Malay *gosi* 'penis', Ambon Malay *gosi* 'male genitalia', Manado Malay *gosi* 'penis'
gutila 'k.o. fish'	*gutila* 'k.o. fish'	
hohati 'fish with rod and line' < *hati* 'fish hook'	*hohate* 'fishing gear, fishing line'	North Moluccan Malay *huhate* 'fish with a pole', Ambon Malay *hohati* 'fishing rod, bamboo rod to which wire is attached for fishing; fish with a fishing rod', Manado Malay *hohati* 'fish with a fishing rod'
kalai ~ kalawai 'barbed spearhead, spear with barbed iron point'	*kalawai* 'fishing spear with multiple barbs'	Ambon Malay *kalawai* 'fish spear, barb'; Manado Malay *kalawai* 'barb'
kiha 'giant taro, *Alocasia macrorrhizos*'	*kiha* 'giant taro, *Alocasia macrorrhizos*'	North Moluccan Malay, Manado Malay *kiha* 'giant taro, *Alocasia macrorrhizos*', Ambon Malay *kaladi kiha* 'kind of taro'
kum-kum 'k.o. dove, *Streptopelia chinensis*'	*kum-kum* 'k.o. pigeon'	North Moluccan Malay *kum-kum* 'brown wood pigeon with long tail', Bacan Malay *kum-kum* 'various k.o. pigeons and doves'
kuso 'cuscus'	*kuso ~ kusu, kus-kus* 'cuscus'	North Moluccan Malay *kuso* 'cuscus', Ambon Malay *kusu* 'phalanger (a kind of marsupial)'†
kusu-kusu 'k.o. grass, *Imperata cylindrica*'	*kusu-kusu* 'kunai grass'	North Moluccan Malay, Ambon Malay *kusu-kusu* 'k.o. grass'

Table 1. *(continued)*

Ternate	Papuan Malay	Other eastern Malay varieties
lasi 'trevally, k.o. fish'	*lasi* 'fish sp.'	Ambon Malay *ikaŋ lasi* 'fish, spotted mackerel'
ligua 'k.o. tree, *Pterocarpus indicus*'	*leŋgua ~ liŋgua* 'k.o. tree, *Pterocarpus indicus*'	Ambon Malay *leŋgua ~ liŋgua* 'k.o. tree, *Pterocarpus indicus*'
lobe 'look for something in the dark with a torch or lamp, fish with lamps', *lobe-lobe* 'lighting devices (such as lamps, torches, torches, etc)'	*lobe* 'torch, night hunt', *balobe* 'search for food on the beach at evening'	Ambon Malay *lobe* 'torch of dried coconut leaves', *balobe* 'go fishing in the tide pools or streams with a sharp object and a torch', Manado Malay *lobe* 'flashlight'
lolaro 'k.o. mangrove tree, *Bruguiera* sp.'	*lolaro* 'k.o. mangrove tree, *Bruguiera* sp.'	
moleo '*Megapodius* sp.'	*maleo* '*Megapodius* sp.'	North Moluccan Malay, Ambon Malay *moleo* '*Megapodius* sp.'
mou 'dumb, mute'	*mou* 'dumb, mute'	North Moluccan Malay, Bacan Malay *mou* 'speechless, mute, speech-impaired'
ŋaju 'outrigger beams, outriggers'	*naju-naju* 'outrigger arm'	Ambon Malay *najoŋ* 'outrigger spar attaching floats to canoe'
ɲare 'reef, sand-bank'	*ɲare* 'coral reef'	Manado Malay *ɲare* 'estuary
ofu 'bee, wasp'	*ofu* 'bee'	North Moluccan Malay *ofu* 'wasp, bee', Manado Malay *ofu* 'bee'
pea-pea 'black-lip pearl oyster, *Pinctada margaritifera*'	*bia pea-pea* 'k.o. shell fish'	North Moluccan Malay *bia pea-pea* 'k.o. shell fish'
poparo 'salt-water fish, barracuda'	*poparo ~ poparu* 'mackerel, fish sp'	Bacan Malay *poparo* 'k.o. large predator fish'
popeda 'sago porridge' ‡	*papeda* 'sago porridge'	North Moluccan Malay *popeda ~ pupeda* 'sago porridge', Ambon Malay *papeda* 'sago porridge'
sabeta 'sago grub'	*sabeta* 'sago grub'	
siafu '*Dioscorea* sp.'	*siafu* '*Dioscorea* sp.'	
soa-soa 'monitor lizard'	*soa-soa* 'lizard'	North Moluccan Malay, Ambon Malay, Manado Malay *soa-soa* 'monitor lizard'

Table 1. *(continued)*

Ternate	Papuan Malay	Other eastern Malay varieties
soma 'drag-net'	*soma* 'large trawl, dragnet'	North Moluccan Malay *soma* 'fishing net', Bacan Malay *soma* 'k.o. fishing net', Manado Malay *soma* 'trawl net, larger than a *kalenda*'
tamako 'axe'	*tamako* 'axe (stone)'	North Moluccan Malay, Manado Malay *tamako* 'axe'
tasa 'bald'	*tasa* 'bald'	North Moluccan Malay, Manado Malay *tasa* 'bald'
tobo 'swim'	*tobo* 'dive'	North Moluccan Malay, Manado Malay *tobo* 'swim
tome-tome '*Flacourtia inermis*'	*tomi-tomi* '*Flacourtia inermis*'	Ambon Malay *tomi-tomi* 'fruit tree', Bacan Malay *tome-tome* 'k.o. tree bearing small, round, edible sour fruit, perhaps, sometimes made into a sugary preserve'
tutaŋ ~ totaŋu 'roofed shelter on a boat'	*totaŋgo* 'shelter on canoe'	
tusa 'cat'	*tusa* 'cat'	North Moluccan Malay, Ambon Malay, Manado Malay *tusa* 'cat'
tum ~ tumu 'dive'	*tum* 'dive'	North Moluccan Malay *tum* 'dive', Bacan Malay *tum* 'sound of splashing or diving'
walu 'bottle gourd, *Lagenaria siceraria*'	*walo-walo* 'bottle gourd'	Bacan Malay *walu* 'various k.o. squash'

* Danckaerts and Wiltens (1623) give Moluccan Malay <*saffarouni*> 'body filth'. Their work combines North Moluccan and Ambon Malay lexemes.
** Standard Malay has *puki mai* as a term of abuse. I regard the expression *cuki mai* in these Malay varieties as a blend in which Ternate *cuki* took the place of standard *puki*, no doubt partly motivated by the obvious formal similarity between the words.
*** Standard Malay has *gara* 'shocking event, commotion, riot' and *gara-gara* 'because', but I regard these as only having chance similarity with the forms presented here.
† The most common form for 'cuscus' in Malay today is *kuskus*. This certainly originates in a reduplication of *kuso ~ kusu*, and indeed *kusu-kusu* is found for 'cuscus' in some 19th century sources. It is likely that the form was reduced to *kus-kus* to differentiate it from the grass, *Imperata cylindrica*, also known as *kusu-kusu* in the Moluccas.
‡ This is an apparent partial reduplication from earlier *peda 'sago'.

As a major centre for the spice trade, Ternate was of enormous importance in the region during the pre-modern and early modern periods. The list in Table 1 includes several lexemes which, on first sight, might be thought to be Malay, but in reality, originate in Ternate. Lexemes for several forms of biota, particularly *Megapodius* and *Phalanger* species, that are unique to eastern Indonesia were adopted from Ternate into Malay, including those in the west of Indonesia. That these originate in Ternate is seen from the fact that they have reconstructive value in the North Halmaheran language family (i.e., Ternate *kuso* < Proto-North Halmahera *kusoC 'cuscus, *Phalanger* spp.', Ternate *moleo* < Proto-North Halmahera *maleuŋ 'maleo, *Megapodius* spp.'). Ternate also acted as a conduit for borrowings from the languages of traders, such as Javanese, Arabic, Portuguese and Spanish, as well as Dutch (Schapper & Zielenbach 2025a, 2025b). This list does not (intentionally) include lexemes that have been borrowed into Ternate and dispersed further throughout the region from that language, especially into Malay varieties, but also into many other languages in the region. Examples of Papuan Malay lexemes of this kind include *gode* 'fat' and *mayana* 'painted nettle'; both originate in Ternate borrowings of (Old) Javanese *gĕḍe* 'largeness, bigness' and *mayana* '*Coleus* spp.'. If all lexemes of this kind were included in Table 1, the list would be substantially longer.

Even to fieldworkers who have spent considerable amounts of time in Papua, numerous Papuan Malay lexemes on this list are likely to be unfamiliar. Most of the Papuan Malay lexemes in Table 1 are not to be found in the available Papuan Malay lists (i.e., Anceaux n.d.; van Velzen 1995; Kluge 2017: 565–621). This reflects a shortcoming in the way Malay varieties tend to be documented; that is, data collection is often carried out in cities and towns, and topics covered in the documentation tend to have a strongly urbane character. We see, however, that a significant proportion of the distinctive vocabulary in Papuan Malay that originates in Ternate refers to items of the natural world, especially animal and plant names, as well as to items of traditional material culture, particularly in the domains of seafaring and hunting. It is thus unsurprising that these lexemes are not captured. Documentation of indigenous languages in Papua is, however, typically conducted in small villages with subsistence economies centred on gardening, hunting and fishing, and accordingly is much more useful in reflecting these distinctive lexical aspects of local Papuan Malay. I was able to extract many of the local Papuan Malay terms for biota from the glosses of lexemes in language documentation corpora. For example, *(pohon/kayu) lolaro* is given as a gloss in Kamholz (2013) for Moor *raindéa* 'tree sp.' and by Jones et al. (1989) for Yawa *panae mot* 'k.o. mangrove tree', thus providing two distinct records and locations in the Cenderawasih Bay area where this Papuan Malay term of Ternate origin is in use.

Papuan Malay has lost some of its distinctive regional vocabulary in relatively recent times and adopted more standard Malay terms. For this reason, it is often useful to look back at older documentation of the lexicon, such as it exists. For instance, while the lexical material of Kluge (2017) only gives *kuciŋ* for 'cat' and *tawoŋ* for 'bee', the Papuan Malay list collected in the first part of the 20th century by J. C. Anceaux gives *kuciŋ* and *tusa* for 'cat' and *ofu* for 'bee'. The Anceaux list is the only record we have of the use of *tusa* and *ofu* in Papuan Malay and harks back to a period in which that language was more heavily influenced by North Moluccan Malay, a variety with very extensive borrowing from Ternate. *Kuciŋ* 'cat' and *tawoŋ* 'bee' reflect more recent influence from standard Malay/Indonesian and are typical of Papuan Malay as it is spoken today. Modernisation and changes in material culture are also undoubtably responsible for changes in the Papuan Malay lexicon, including its move away from earlier vocabulary originating in Ternate. Take, for example, *gau* 'piece of the young leaf of the areng palm used as cigarette paper', a Papuan Malay lexeme recorded in the Anceaux (no date) list. Nowadays the majority of tobacco consumed in Indonesia is in the form of factory-produced cigarettes. The lexeme *gau* is, again here, only recorded in languages in the region where there is older documentation, such as Ternate and Manado Malay, or where the lexeme has, through semantic shift, come to refer to the modern cigarette, as is apparently the case in many Gorontalo-Mongondow languages (Usup 1986: 310) after they borrowed the item from Manado Malay.

Differences in the number and identity of Ternate loans in Malay varieties around Halmahera point to different historical forces at play, that is, different time-depths in the formation of the Malay varieties as well as, potentially, different pathways by which Ternate lexemes entered the Malay varieties. The fact that some of the Papuan Malay lexemes originating in Ternate are not known in other Malay varieties in the region suggests that some of the lexical influence from Ternate spread independent of Malay, that is, not necessarily from North Moluccan Malay into Papuan Malay. However, this may simply be an artefact of the limited lexical documentation that we have for Malay varieties in certain domains, as observed above. There are also certain features of Papuan Malay that indicate that the spread of Ternate lexemes was mediated by Malay varieties. Certain phonological adaptations of Ternate loans are shared across the same lexemes in Malay varieties in the region including Papuan Malay (e.g., such as sporadic fortition of Ternate medial *g* to *ŋg*, for instance, in Ternate *ligua* '*Pterocarpus indicus*' > Ambon/Papuan Malay *leŋgua* ~ *liŋgua*). Certain compounds consisting of a Ternate lexeme and a Malay lexeme are shared across Malay varieties (e.g., North Moluccan/Ambon/Papuan Malay *hati gaba* 'lungs' < Malay *hati* 'liver' and Ternate *gaba(-gaba)* 'lungs'; North Moluccan/Ambon/Papuan Malay *poro bibi* 'pot-belly pufferfish' < North Moluc-

can/Ambon Malay *poro* 'belly' (cf. Standard Malay *pərut*) and Ternate *bibi* 'pot-bellied, pufferfish'). Finally, certain semantic innovations in Ternate lexemes seem to be shared across Malay varieties (e.g., North Moluccan/Ambon/Papuan Malay *gosi* ~ *gusi* 'penis' < Ternate *gosi* 'egg, testicle').

Additional evidence for Papuan Malay's close links to other Malay varieties in the region comes from the fact that specific lexemes are also shared amongst them. For example, Papuan Malay *molo* 'dive' is also found in North Moluccan Malay, Ambon Malay and Banda Malay, as well as many indigenous languages of the Central and Southern Moluccas. Papuan Malay *kelemarin* 'yesterday', mentioned by Donohue (2003) and also used by Doble (1960) in the glosses of her Mee dictionary (also illustrating the usefulness of examining these for insights into older Papuan Malay lexical forms), is found in North Moluccan, Ambon and Manado Malay and reflects an archaic form still extant in Classical Malay *kelmarin* (< Proto-Malayic *ke-le[h,ʔ]em hari-an, (Adelaar 1992: 113)).[1] There is one form that suggests a closer relationship of Papuan Malay to North Moluccan Malay than to other varieties in the region: Papuan Malay and North Moluccan Malay have *pepe* 'vagina', while Ambon and Banda Malay have *popo* 'vagina'. I suspect that both of these originate in a figurative extension of Ternate *popo* 'split, stumped' that occurred when it was loaned into Malay (cf. the semantic shift of Ternate *gosi* 'egg, testicle' to mean 'penis' when loaned into local Malay varieties). I hypothesise that the form *popo* in Ambon and Banda Malay is the original form of the lexeme 'vagina' in North Moluccan Malay, and that the form *pepe* represents a phonological tampering of *popo* motivated by word tabooing, as is common cross-linguistically (cf., *shit* > *shoot, fuck* > *fudge*). In this scenario, *pepe* is a later innovation in North Moluccan Malay that then spread to Papuan Malay. This also places the formation of Papuan Malay later than that of other varieties such as Ambon Malay.

There is some very limited influence on Papuan Malay from Ambon Malay, but it concerns only a small number of lexemes that are highly circumscribed geographically. For instance, in the region in and around the city of Sorong, *niri* is used for 'bee, wasp'; it comes from Ambon Malay *niri* 'bee, wasp' (from indigenous Ambon languages *niri* 'tooth', but ultimately from PMP *ba|**ŋ**elih** 'tooth'), but is unattested elsewhere in Papua.[2] Similarly, *usi* 'older sister' in the Papuan Malay of Serui likely originates in Ambon Malay (van Velzen 1993: 333), but again

1. Most recently, Gil (2022: 495) notes *kelemarin* in the Papuan Malay of Manokwari.

2. It is not always possible to determine the route by which specific lexemes entered the Papuan Malay of Sorong. According to the information available to me, *ikan gaca* (< Ternate *gaca* 'red snapper') is only used in Sorong. It is unclear whether the term was borrowed through North Moluccan Malay or Ambon Malay, both of which attest the item.

it is not attested elsewhere in Papuan Malay to my knowledge.[3] That lexemes with specifically Ambon Malay origins have such restricted numbers and ranges in Papuan Malay indicates that influence from Ambon Malay was late and of very limited significance, particularly when compared to that issuing from North Moluccan Malay, as van Velzen (1993: 315) points out (*pace* Roosman 1982).

3. The role of a Malay trade pidgin in the formation of Papuan Malay

We saw in Section 1 that the emergence of Papuan Malay has been repeatedly connected in the literature with the idea that there was longstanding use of Malay in trade along the north coast of New Guinea. This idea, however, has been challenged by Seiler (1985), *inter alia*; very little evidence exists to suggest that there was much Malay competence before the 20th century. Rather, local lingua francas seem to have been the primary vehicle for trade from Raja Ampat to the Mamberamo River: Biak played the most extensive role being used widely in inter-group communication in the coastal regions through Raja Ampat over the Bird's Head and into Cenderawasih Bay and along the north coast of Yapen, while Wamesa was the widely known language in southern Cenderawasih Bay and along the south coast of Yapen. Van Asperen (1936) makes clear that these languages were still the primary vehicles for trade in Cenderawasih Bay well into the 20th century:

> Biak and Wamesa are understood by large swathes of people. Biak is traditionally the language of general use between ethnic groups, the members of which speak different languages. The majority of traders learn Biak or Wamesa because the Biak and Wamesa-speaking peoples do not show any inclination to switch over to Malay.[4] (van Asperen 1936: 75–76)

See Schapper & van Schie (to appear) for a comprehensive overview of historical sources relating to the use of Biak and Wamesa as lingua francas.

---

3. *Usi* 'older sister' is not unique to Ambon Malay; it is also found in Kupang Malay and the Timor region. Some relational terms in Ambon Malay were introduced from Timor by the Portuguese, e.g., *fetor* 'trader, ruler'. As such, it is not entirely clear that *usi* is even original to Ambon Malay.

4. Original: "Noemfoorsch en Windesisch worden er door groote volksgroepen verstaan. Het Noemfoorsch is van oudsher de omgangstaal tusschen stammen, wier dialecten afwijkend waren. De meeste handelaren leeren het Noemfoorsch of Windèsisch aan, omdat de Noemfoorsch en Windèsisch sprekende stammen niet de neiging vertoonen om het Maleisch over te nemen."

East of the Mamberamo River, however, little is known about historical language use. In this section, I show that there is evidence for contact with traders speaking Malay along the north coast of New Guinea, beyond Cenderawasih Bay. I suggest that while this contact might have resulted in some competence in Malay for particular local individuals, it does not seem to have resulted in any lexical contribution to Papuan Malay as is spoken today.

Historical documents dealing with the north coast of New Guinea beyond Cenderawasih Bay are very limited, and notes on language use few and far between. Still, a picture of relative isolation emerges from the historical accounts that do exist. The Etna expedition reached Humboldt Bay in 1858 and reported that the Papuans living there did not know any Malay and had no contact with the outside world, writing:

> As a matter of fact, both of these peoples [i.e., the Malays and the Tidorese mentioned in the preceding text] never reached as far as Humboldt Bay with their sea voyages, and we are not aware of any other indigenous sea-faring peoples who have reached the bay or any of the other preceding or nearby capes that we have named.[5] (van der Goes 1862: 82)

From a visit in 1871 to multiple places along the north coast, including Humboldt Bay, it was reported to be very difficult to communicate with the locals in Malay (Robidé van der Aa 1879: 109–116, 127–132). The missionary Bink travelled to Humboldt Bay in 1893, but did not note any Malay competence explicitly in his account; he recognised the existence of trade in massoi, mother-of-pearl, and sea cucumber (Bink 1894). In 1903, despite a camp of traders in the area, Wichmann (1917) noted that just one local person in Humboldt Bay could act as interpreter on account of having spent time in Ternate. Finally, during a visit to Jamna Island in 1911, van Hasselt (1926: 134) reports that several Papuans could speak Malay due to regular contact with traders. In short, the available historical records suggest that knowledge of Malay was almost non-existent before the beginning of the 20th century and only was developed by a relatively small number of people in contact with traders who began arriving at the end of the 19th century.

Who were the traders who began the introduction of Malay on the northeast coast? Only Bink (1894) is explicit about the identity of the visitors to the north coast of New Guinea, noting that there were bird hunters from Ternate in the Humboldt Bay area. Wichmann (1917) refers simply to "Malay" traders, but this term is regularly applied to any peoples from to the west of New Guinea speaking

5. Original: "Trouwens beide deze volken hebben hunne zeetogten nimmer tot de Humboldtsbaai uitgestrekt, en het is ons niet bekend kunnen worden, dat andere inlandsche zeevarende volken gezegde baai hebben bezocht, of aan een der voor of nabij haar gelegen kapen den voorzegden naam hebben gegeven."

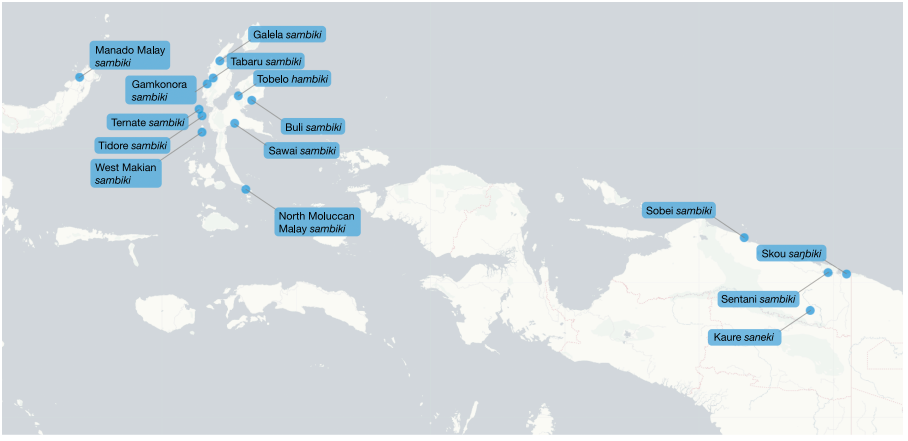
a variety of Malay; it does not refer to ethnic Malays from western Indonesia.[6] Loanwords into the indigenous languages of the north coast allow us to be more precise in identifying the first contacts of the region. Consider the distribution of loanwords referring to introduced items in the maps below.

The mapped data presents a clear picture of contact between peoples of the north coast of New Guinea and of Halmahera, specifically speakers of North Moluccan Malay. Portuguese *milho* 'maize', referring to an originally Latin American crop, is uniquely loaned into North Moluccan Malay, and from there into Manado Malay. This provides the only source for the borrowing of this item into the languages of the north coast of New Guinea, barring Portuguese speakers themselves (Map 2). Ternate *sambiki* 'gourd', apparently referring originally to gourds of the genus *Lagenaria* introduced from Africa, is loaned into a range of North Moluccan languages including North Moluccan and Manado Malay, but then also appears in the languages of the north coast of New Guinea (Map 3). Similarly, Ternate *tusa* 'cat', referring to a creature first thought to have reached the region on the boats of European traders, is loaned into a range of languages in the Moluccas including regional Malay varieties, but in New Guinea exclusively appears on the north coast of New Guinea (Map 4).



Map 2. Loans of Portuguese *milho* 'maize'

6. There is a pervasive and pernicious practice in the literature on the history of New Guinea whereby almost anything that comes from Indonesia is referred to as "Malay". For example, in a recent article, Douglas (2024) discusses indigenous knowledge of New Guinea as "Malay", even though the knowledge referred to is clearly not that of Malays, but rather (mostly) from people of the Moluccas, specifically what would today be referred to as the North Moluccas. Linguists suffer less from this disease than scholars of other disciplines, but it is surprising how common language use and ethnicity are conflated even by linguists, when it comes to Malay.



Map 3. Loans of Ternate *sambiki* 'k.o. gourd, melon'[7]



Map 4. Loans of Ternate *tusa* 'cat'

The strikingly gapped distribution of these loanwords, whereby they are in evidence in the North Moluccas and on the north coast of New Guinea, but are absent in the languages of Raja Ampat, the Bird's Head and Cenderawasih Bay, points to speakers of North Moluccan Malay travelling to the north coast of New Guinea and having direct contact with the local peoples, without the mediation of traditional trading groups like the Biak from further west. In the case of the intro-

7. The Ternate form is almost certainly related to the archaic Malay *tambikai*. The initial /s/ of the Moluccan forms means that we can be sure of the Ternatean origin of the forms along the north coast of New Guinea.

duction of cats to the Humboldt Bay area, cross-referencing with the historical record confirms the loanword inference made here — with the additional of some astonishing detail: Bink (1902: 65) remarks that the first cat arrived from Ternate in August 1893 and that the indigenous peoples did not have a name for the creature.

Not all the loanwords on the north coast of New Guinea that originate in the North Moluccas have this gapped distribution, but there is often still reason to see speakers of North Moluccan Malay as the immediate source. Consider the loans of Ternate *gomala* ~ *gumala* 'fishhook' in Map 5. This lexeme appears in several languages of Cenderawasih Bay, but in each case it has the phonological adaption of *l* > *r*, as is typical of Biak-mediated loans (Schapper & van Schie to appear). The fact that this adaption is missing in Mekwei *gumala* 'fishhook' points to the loan being directly loaned from Moluccans and not via a language in Cenderawasih Bay. This analysis of the Mekwei form is possible because Mekwei has a phonemic contrast of /l/ and /r/, whereas Berik and Isirawa have only a single liquid /r/. Additionally, all forms of this item in Cenderawasih Bay have a penultimate back vowel, indicating that they are borrowed from Tidore *gomola*. The north coast New Guinea languages discussed here have a penultimate /a/ vowel, which points to Ternate *gomala* as their ultimate source. Not all loans from the North Moluccas on the north coast of New Guinea have tell-tale sound changes that allow us to positively identify the immediate source. Loans of Ternate *fofoki* ~ *fokifoki* 'eggplant, aubergine' are one such case (Map 6). What is significant here, however, is that there is no positive evidence to suggest that Biak or another language acted as an intermediary for this loan.



Map 5. Loans of Ternate *gomala* ~ *gumala* 'fishhook'



Map 6. Loans of Ternate *fofoki* ~ *fokifoki* 'eggplant, aubergine'

A string of isolated loanwords in individual languages of the north coast of New Guinea provide additional support for the idea of North Moluccan contact with people along the north coast of New Guinea. These include: Sentani *doroku* 'bridge' < Ternate *dodoku* 'bridge' (cf. Bacan Malay, Manado Malay *dodoku* 'bridge'), Tarpia *popoji* 'pocket, bottle' < Ternate, North Moluccan Malay *popoje* 'pocket, pouch', Sobei *yakis* 'monkey' < North Moluccan Malay *yakis* 'monkey', Tarpia *baramamsa* 'lemongrass' < Tidore *baraŋmakusu* 'lemongrass' (cf. Manado Malay *bramakusu* 'lemongrass'), Tarpia *doda* 'anchor' < Ternate *doda* 'anchor'. Only some of these loans are attested in trade languages of the north coast of New Guinea; they cannot all be argued to have been dispersed from the North Moluccas through such intermediaries.

Taken together, the loanwords presented in this section establish that there was indeed trade contact between indigenous groups on the north coast of New Guinea and Moluccans speaking a variety of North Moluccan Malay with influences from important languages like Ternate and Tidore. However, it is very difficult to see from the evidence that this trade and the language contact arising from it was foundational to the emergence of Papuan Malay. The loanwords that we find along the north coast of New Guinea are almost uniformly not part of what is known as Papuan Malay, either today or as it has been historically recorded. If Papuan Malay initially developed through trade contacts with Moluccans as has been claimed, then we would expect that at least some of the lexemes which were loaned at the time of this contact would not only be present in the local languages of north coast New Guinea, but also reflected in Papuan Malay as it became established.

4. Conclusion and discussion

This chapter has looked at aspects of the lexical history of Malay in the north-western part of New Guinea with the aim of clarifying Malay's role as a contact language in New Guinea and the emergence of Papuan Malay. It was shown that Papuan Malay has numerous lexemes that originate in the Ternate language. This places it firmly within the network of neighbouring Malay varieties in eastern Indonesia which show similar lexical influence from Ternate. Some lexical items link Papuan Malay most closely to North Moluccan Malay, while others present in the areas of Sorong and Serui show influence from Ambon. These facts suggest that Papuan Malay has almost certainly been formed in part through contact with speakers of North Moluccan and Ambon Malay. It is extremely difficult to argue on the available linguistic evidence that Papuan Malay emerged from long-standing trade networks that connected Papuan peoples with the Moluccas. The historical record also makes clear that Biak and Wamesa were used in intergroup trade from Raja Ampat to Cenderawasih Bay. To the east of Cenderawasih Bay, there appears to have been very little contact with the outside world until the very end of the 19th century. The loanwords on the north coast of New Guinea that resulted from contact with Moluccans are not the same as those found in Papuan Malay today. That contact cannot, therefore, be argued to have been instrumental in the initial formation of Papuan Malay.

Where, then, did Papuan Malay come from? My view broadly agrees with that of Kluge (2017: 46): the most likely scenario is that Papuan Malay arose through formal institutions that were established from the beginning of the 20th century. The Dutch authorities began to enforce the use of Malay in education, administration and religious settings in the western half of New Guinea at this time. This meant recruiting personnel from the Malay-language school system already established in the Moluccas to work as administrators, teachers and in other professions (Collins 1998: 64). These people spoke varieties of Malay that had been historically influenced by Ternate. This lexical influence from Ternate would then have carried on into Papuan Malay as it formed, reinforced by whatever limited, Ternate-influence Malay competence was already in place among local elites. Such a scenario involving diverse inputs explains why Papuan Malay has a unique constellation of Ternate loans, distinct from that found in other Malay varieties in the region. It is for historians to look at the policies and practices of the Dutch colonial government and of the Christian missions at these times to establish the process of introducing Malay and Malay-speaking peoples into Papua.

Linguists working in Papua, particularly those concerned with Malay varieties, have shown little inclination to document lexicon, following the Bloombergian tradition of (implicitly) regarding the lexicon as a mere appendix to the grammar of a

language. The documentation of the lexicon and lexical diversity — not only in the different towns of Papua, but also in remote villages — is in fact crucial for understanding the history of contact languages in the region. In this chapter we saw that dictionaries of indigenous languages around the Bird's Head that included Malay glosses revealed many more lexical items of Ternate origins than was immediately obvious from documentation of Papuan Malay. This speaks to the limitations of documentation in urban settings where key semantic domains — particularly, plants and animals — rarely feature.

There is still a rich world of Papuan Malay to document, particularly in the domain of the lexicon. Papuan Malay is developing its own characteristics and becoming more distinctly Papuan, with influences from local substrates continuing to percolate up. In the Papuan Malay of Manokwari, *keladi* 'taro' and *maŋga* 'mango' are figuratively extended to also mean 'penis' and 'breast', respectively (Nando Saragih pers. comm.). This development is consistent with common lexification patterns in Papuan languages whereby the human body is in an extended parallel with (parts of) plants. In the Papuan Malay of Serui, *bia* 'shell, shellfish', originally from Ternate, has taken on the additional meaning of 'vagina' (van Velzen 1995: 331), paralleling the colexification patterns present in the local languages of Yapen (e.g., Ambai *firai* 'shell, vagina'). We also see many other unique coinages in Papuan Malay, such as *papeda laut* [sago.porridge sea] 'jelly-fish' and *peɲu belimbiŋ* [turtle star.fruit] 'leatherback turtle', that reflect the penchant of Papuan languages for using similarity-driven colexifications in referring to biota (Schapper, to appear).

My own work has taken me only very recently to Papua, unintentionally, following in Nikolaus' footsteps to Manokwari. Apparently perceiving my lack of intrepidity and general ignorance of the city, Nikolaus insisted on taking me on what I can only describe as an utterly hair-raising motorcycle trip to see the historical sights of Manokwari. The history lesson was much appreciated, but next time, I might just walk. Or, maybe, I will just stay sitting on the plastic chairs with the cats.

Funding

Research for this chapter was supported by the OUTOFPAPUA project funded by the European Research Council (grant agreement no. 848532).

Acknowledgements

Many thanks to Anne van Schie who located some of the historical sources discussed in this chapter and to David Gil and Sander Adelaar who commented on earlier versions. All errors are my own.

References

Adelaar, K. Alexander. 1992. *Proto-Malayic: The Reconstruction of Its Phonology and Parts of Its Lexicon and Morphology*. Canberra: Pacific Linguistics.

Adelaar, K. Alexander, and David J. Prentice. 1996. "Malay: Its History, Role and Spread." In *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas*, ed. by Stephen A. Wurm, Peter Mühlhäusler, and Darrell T. Tryon, 673–693. Berlin: De Gruyter Mouton.

Anceaux, J.C. no date. Nieuwguinees Maleis — Nederlands. *Leiden University Special Collections, J.C. Anceaux Collection, Or.* 615, Folder 88.

van Asperen, Lubbert. 1936. *Zending en Zendingsonderwijs op ederlandsch-Nieuw Guinee*. Leiden: M. Dubbeldeman.

Bink, G.L. 1894. "Tocht van den zendeling Bink naar de Humboldts-baai." *Tijdschrift van het Aardrijkskundig Genootschap* 11: 325–332.

Bink, G.L. 1902. "Lijst van woorden opgeteekend uit den mond der Karau Jotafa, bewoners des Humboldts-Baai." *Tijdschrift van het Bataviaasch Genootschap* 45: 59–89.

Collins, James T. 1980. *Ambonese Malay and Creolization Theory*. Kuala Lumpur: Dewan Bahasa dan Pustaka and Kementarian Pelajaran Malaysia.

Danckaerts, Sebastian, and Casper Wiltens. 1623. *Vocabularium ofte Woordenboeck, nae ordre van den Alphabeth in 't Duytsch en Maleys*. 's Gravenhage: Ordinaris Drukkers vande Hog.

Doble, Marion. 1960. *Kapauku-Malayan-Dutch-English dictionary*. The Hague: Martinus Nijhoff.

Donohue, Mark. 2003. "Papuan Malay." Unpublished manuscript, Singapore.

Douglas, Bronwen. 2024. "'Very Beautiful Land': Malay Knowledge, Spanish Voyages, and Indigenous Presence in Iberian Mapping of New Guinea." *The Journal of Pacific History* 59: 1–45.

Gil, David. 2022. "Number in Indonesian." In *Number in the World's Languages*, ed. by Paolo Acquaviva, and Michael Daniel, 457–503. Berlin: De Gruyter Mouton.

Gil, David. 2024. "Recent Contact Features in Northwest Papuan Malay." In *Language Contact Between Malay and Indigenous Languages*, ed. by Hiroki Nomoto, and Asako Shiohara, 53–116. Tokyo: NUSA.

van der Goes, H.D.A. 1862. *Nieuw Guinea, ethnographisch en natuurkundig onderzocht en beschreven in 1858*. Amsterdam: Frederik Muller.

van Hasselt, Frans J.F. 1926. *In het land van de Papoea's*. Utrecht: Kemink and Zoon.

Himmelmann, Nikolaus P., and Sonja Riesberg. 2012–2016. "Papuan Malay." In *Collection CELD Papua*. Nijmegen: The Language Archive. https://hdl.handle.net/1839/00-0000-0000-0019-FF78-5

Jones, Linda K., Yohannes Paai, and Zeth Paai. 1989. *Mo mona nanentabo ranugan — Perbendaharaan kata Bahasa Yawa — Yawa Vocabulary*. Cenderawasih University and Summer Institute of Linguistics.

Kamholz, David. 2013. "Toolbox lexicon of Moor." http://lexifier.lautgesetz.com

Kluge, Angela. 2014. A Grammar of Papuan Malay. PhD thesis, Leiden University.

Kluge, Angela. 2017. *A Grammar of Papuan Malay*. Berlin: Language Science Press.

Paauw, Scott H. 2009. The Malay Contact Varieties of Eastern Indonesia: A Typological Comparison. PhD thesis, State University of New York.

Robidé van der Aa, Pieter J. B. C. 1879. *Reizen naar Nederlandsch Nieuw-Guinea ondernomen op last der regeering van Nederlandsch-Indië in de jaren 1871, 1872, 1875-1876 door de heeren P. van der Crab en J. E. Teysmann, J. G. Coorengel en A. J. Langeveldt van Hemert en P. Swaan met geschied- en aardrijkskundige toelichtingen*. 's Gravenhage: Martinus Nijhoff.

Roosman, Raden S. 1982. "Pidgin Malay as Spoken in Irian Jaya." *The Indonesian Quarterly* X (2): 95–104.

Rowley, Charles D. 1972. *The New Guinea Villager: A Retrospect from 1964*. Melbourne: F. W. Cheshire.

Saragih, Chrisma Fernando. 2012. The Practical Use of Person Reference in Papuan Malay. Master's thesis, Radboud University.

Schapper, Antoinette. To appear. "Lexical Typology, with Special Reference to Colexification." In *The Oxford Guide to Papuan languages*, ed. by Nick Evans, and Sebastian Fedden. Oxford: Oxford University Press.

Schapper, Antoinette, and Maria Zielenbach. 2025a. "Dutch Loanwords in Ternate and Their Circulation in the Moluccas." In *Dutch and Contact Linguistics*, ed. by Christopher Joby, and Nicoline van der Sijs, 392–426. Amsterdam: John Benjamins.

Schapper, Antoinette, and Maria Zielenbach. 2025b. "Reframing Pre-modern Language Contact Through Trade in Eastern Indonesia: Javanese Linguistic Influence in the Moluccas." *Journal of the Southeast Asian Linguistic Society* 18 (2): 1–23.

Schapper, Antoinette, and Anne van Schie. To appear. "Lingua Francas of the SHWNG Region: Untangling the Historical Dynamics of Biak and Wamesa Loanwords." In *New Perspectives on the History of SHWNG Languages*, ed. by Laura Arnold, and Antoinette Schapper.

Seiler, Walter. 1985. "The Malay Language of New Guinea." In *Papers in Pidgin and Creole Linguistics*, ed. by Stephen A. Wurm, 143–153. Canberra: Pacific Linguistics.

Usup, Hunggu Tadjuddin. 1986. Rekonstruksi proto-bahasa Gorontalo-Mongondow. PhD thesis, Universitas Indonesia.

van Velzen, Paul. 1995. "Some Notes on the Variety of Malay Used in Serui and Vicinity." In *Tales From a Concave World: Liber amicorum Bert Voorhoeve*, ed. by Connie Baak, Mary Bakker, and Dick van der Meij, 311–343. Leiden: Department of Languages and Cultures of Southeast Asia and Oceania, Leiden University.

Wichmann, Arthur. 1917. *Bericht über eine im Jahre 1903 ausgeführte Reise nach Neu-Guinea* [*Nova Guinea: Résultats des expéditions scientifiques à la Nouvelle Guinée en 1903 sous les auspices de Arthur Wichmann, Vol. 4*]. Leiden: E. J. Brill.

CHAPTER 31

On the origins of Javanese negators

Alexander Adelaar
Palacký University Olomouc | The University of Melbourne

This paper investigates the origins of Old and Modern Javanese negators. Modern Javanese distinguishes four negators.

- Predicative *ora* ('no(t)') derives from Old Javanese *tan+wwara* 'there is not' and its high register counterpart *botən* from *tan+wontən*.
- Prohibitive *aja* ('don't') and its Old Javanese predecessor *haywa* must derive from Old Javanese *hayu* 'beauty, goodness, rightness; beautiful, good, right' + subjunctive *-a: hayu-a* was used in desiderative and hortative phrases ('it would be good if…'), and, with *tan*, in prohibitives. After loss of *tan, hayu-a* became lexicalised as *haywa*, a prohibitive. Its high counterpart *sampun* is basically an aspect marker ('already'), which acquired a cessative and prohibitive function.
- Contrastive *dudu* ('no [+noun]; not this [but another]') originally meant 'other'.
- The origin of Phasal *duruŋ* 'not yet' needs further investigation.

Keywords: negators, Old Javanese, Modern Javanese, historical linguistics

1. Introduction

In this chapter I investigate the origins of Javanese negators, concentrating on their lexical development from Old Javanese to Modern (present-day) Javanese, a standard Javanese form deriving from the court languages of Yogyakarta and Solo in Central Java. In doing so I use of a combination of various historical linguistic methods including internal reconstruction, the comparative method, and philology. I also make use of Harris and Campbell's (1995) re-analysis theory, Mańczak's (1980) theory regarding irregular sound change due to frequency (henceforth ISCDF), and the Negative Existential Cycle model as developed by Croft (1991) and elaborated on by Veselinova (2014, 2016) and Veselinova and Hamari (2022).

https://doi.org/10.1075/slcs.240.31ade
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

I discuss four types of negators that occur most frequently in Modern Javanese, to wit predicative negators, prohibitives, phasal negators, and nominal/contrastive negators. (The latter negate nouns or occur in contrastive contexts; see Section 4). As shown in Section 3, these negators also vary according to social speech register. I address this variation in my discussion, and I also compare Modern Javanese negators with their Old Javanese counterparts wherever these exist.

This chapter is organised as follows. Section 2 clarifies the historical relationship between Old Javanese and present-day Javanese. Section 3 gives a short explanation of speech registers. Section 4 is an overview of the forms and meanings of the negators under discussion, and Section 5 is a justification for their selection. Section 6 discusses predicative negators: Section 6.1 focuses on the Ngoko (i.e. "low Javanese") predicative negator *ora*, including the application of the Negative Existential Cycle and an assessment of its assumed duration; Section 6.2 focuses on its Krama (i.e. "high Javanese") counterpart *botən*. Prohibitive markers[1] are treated in Section 7: Section 7.1 deals with Modern Javanese *aja* and Old Javanese *haywa*, Section 7.2 gives an account of Blust et al.'s (2023; henceforth BTS) perspective on the origin of *aja*, and Section 7.3 treats the Modern Javanese Krama form *sampun*. Phasal negators are discussed in Section 8, and nominal/contrastive negators in Section 9. Concluding remarks follow in Section 10.

In the following pages, the term "formal" usually refers to the shape ("form") of a word (rather than to its use in formal versus informal settings). Words and sentences in quoted languages other than English or proto-languages are in italics; they are represented in the way they were elicited or found in the sources, except for the following sounds, which are represented phonemically:

- schwa (often spelled as <e> or <ĕ> in Modern Javanese): ə
- retroflex voiced stop (spelled <dh> or <ḍ> in Modern Javanese; it is also spelled <ḍ> in Old Javanese and in Dempwolff's (1938) Proto Malayo-Polynesian (henceforth PMP) language material): ɖ
- retroflex voiceless stop (spelled <th> or <ṭ> in Modern Javanese and <ṭ> in Old Javanese and in Dempwolff's (1938) PMP language material): ʈ
- velar nasal (spelled <ng> in Modern Javanese): ŋ
- the palatal nasal (spelled <ny> Modern Javanese): ñ

Note also that Modern Javanese distinguishes between close-mid [e] and open-mid [ɛ], which are orthographically represented as <é> and <è> respectively.

Unless mentioned otherwise, I obtained Old Javanese research data forming the basis of this study from Zoetmulder's (1983) grammar of the Old Javanese Adi-

1. I prefer "prohibitive marker" (or simply "prohibitive") to "prohibitive negator" as the latter has a pleonastic ring to it.

parwa text (estimated to be from the 10th century CE), his survey of Old Javanese literature (Zoetmulder 1974), and his two-volume Old Javanese — English dictionary (Zoetmulder 1982).

I obtained Modern Javanese data through consultation of Uhlenbeck's study of Javanese morphology (1978), Vander Klok's introduction to the languages of Java (2024), dictionaries from Gericke et al. (1901), Pigeaud (1938), and Robson and Wibisono (2002), and from language elicitation and verification with Elisabeth Riharti and Susi Moeimam, native speakers of Javanese. Other sources that I have consulted for these and other languages will be indicated in the text. My own competence in Javanese is based on three years of language training at Leiden University (including a one-year course of Old Javanese) and a four month stay in Central Java (1972–1976).

2. Old Javanese and Modern Javanese

Old Javanese and Modern Javanese are languages in their own right. They are originally historiolects of the same language, although they are not related through a direct evolutionary line (Adelaar 2011: 341–2).[2] Moreover, the periods in which these historiolects were spoken overlap to some extent. As a written language, Old Javanese was in use from the early 9th till the 15th century CE in Java, and it continues to be used as a liturgical language in Bali. Modern Javanese appeared as a written language in the 16th century. The variety cultivated at the courts of both Solo (or Surakarta) and Yogyakarta in Central Java became the basis of Standard Modern Javanese. A third historiolect, Middle Javanese, belongs to a period intermediate between Old and Modern Javanese. Texts in it are from the 14th century onwards (Ogloblin 2005: 590).

2. The notion of historiolects not being related in a direct evolutionary line is in line with what is known more generally about how historiolects of a "same language" relate to one another. As a rule, the recent historiolect has not evolved directly from the old historiolect but from a dialect which was closely related to the latter and co-existed with it but for some reason lost its prominence. A well-known example are the modern Romance languages deriving from Vulgar Latin rather than from Classical Latin.

3. Speech registers

Modern Javanese has several social speech registers, and negators are differentiated accordingly. The two main registers are called "Ngoko" ("low" Javanese) and "Krama" ("high" Javanese). A third register, "Madya", takes an intermediate position between Ngoko and Krama and is less prominent in both form and function. Of the three registers, Ngoko is the least marked: form a historical linguistical perspective, it represents the most authentic form of Javanese. The distinction of registers can be traced to the 16th century CE. It has reached its full development in Modern Javanese but was nearly absent in Old Javanese (Ogloblin 2005: 591). "Kramafication" (the derivation of Krama forms from Ngoko ones, Uhlenbeck 1978: 288) affects basic lexicon and some morphosyntax. In the lexicon it takes various forms and (1) may involve vowel change, e.g. (Ngoko) *dudu* vs. (Krama) *dédé*, the nominal/contrastive negator (see Section 4); *duruŋ* vs. *dèrèŋ* [dɜrɜŋ] 'not yet'; *lumah* vs. *lémah* 'lie on the ground; slow, lazy', or *tamu* vs. *tami* 'guest'; *èlmu* vs. *èlmi* '(occult) science'; (2) syllable change, e.g. *kira* vs. *kintən* 'thought'; *soré* vs. *sontən* 'evening'; *kari* vs. *kantun* 'stay/be left behind'; or *warta* vs. *wartos* 'news, message'; *rəti* vs. *rətos* 'well looked after'; *di-anti-anti* vs. *dipun-antos-antos* 'waited for in anticipation'; (3) lexical replacement, e.g. *loro* vs. *kalih* 'two'; *təlu* vs. *tiga* 'three'; *lima* vs. *gaŋsal* 'five'; *bali* vs. *waŋsul* 'go back'; *aku* vs. *kula* '1st person'; *sapa* vs. *sintən* 'who'; *baé* vs. *kémawon* 'only, just'. As seen in some of the listed examples, it also affects some grammatical words and morphemes (compare the above Ngoko undergoer voice prefix *di-* and its Krama counterpart *dipun-* in *dipun-antos-antos*). See Uhlenbeck (1978) and Vander Klok (2024) for further details.

Although *-ntən* substitution is primarily associated with speech registers and the Kramafication process in Modern Javanese, it is worth noticing that it also occurs in some Old Javanese word pairs:

(1)

Ngoko	Krama
jalu 'male, man, husband'	*jantən*
sāri 'essence, quintessence; pollen, flower'	*santən*
māri, mari 'to cease'	*mantən*
soré 'evening'	*sontən*

Kramafication is not expected to have already happened in Old Javanese. Although this requires further investigation, one might speculate that *-ntən* substitution was an early onset to Kramafication in this historiolect. Adaptations like *-ntən* substitution, which are nowadays associated with Kramafication, may already have been used for the creation of honorific words even before the actual speech registers came into being.

4. The negators: Forms and meanings

In this chapter I focus on the following negators:

1.	Predicative negator	'not'
2.	Prohibitive marker	'don't'
3.	Phasal negator	'not yet' (exceptionally also 'no longer')
4.	Nominal/ contrastive negator	'1. no (+ noun); 2. not [this] (but another noun/verb/ adjective/prepositional phrase)'

"Predicative negation" is the basic means to negate declarative main clauses. In Javanese, these clauses have verbal, adjectival, or nominal predicates and make use of a "predicative negator" *ora* as the basic negating element.[3] The above labels are more accurate than the more common labels "standard negation" and "standard negator".[4] In what follows, I will use them for Javanese, and I will reserve "standard negation" and "standard negator" in reference to languages in which the negator of declarative main clauses is only applied to verbs.

"Prohibitive negation involves calling the hearer(s) not to realize some state of affairs" (van der Auwera et al. 2024: 883). The marker of such a negation (usually occurring at the head of a clause) is a prohibitive marker.

Phasal negators correspond to *not yet* and *no longer* in English and are paradigmatically related to their positive counterparts *already* and *still* (Vander Klok 2024: 875). In Modern Javanese, *duruŋ* (Krama *dèrèŋ*) 'not yet' is the only phasal negator.

The Modern Javanese nominal/contrastive negator (Ngoko *dudu*, Krama *dédé/sanès*, Madya *dédé*) negates nominal phrases. It also negates a (nominal or other) clause indicating that its content is incorrect and in contrast with a following clause which is sometimes left implicit, and the content of which is considered more accurate ("not A [but B] is the case").[5]

---

3. This word is pronounced [ora] and does not follow the general rule in Javanese to round final *-a* to [ɔ]. Note that BTS write <orā> with a macron over the *a*. This is unconventional and does not suggest a long *a*.

4. "Standard negation" only refers to the negation of declarative verbal main clauses (see Payne 1985 and Miestamo 2007: 553), a definition too narrow to be applied to Javanese, as it excludes adjectival and nominal main clauses. The labels "predicative negation" (and, by extension, "predicative negator") were suggested to me by Jozina Vander Klok (written communication July 2025).

5. Kroeger's (2014) analysis of Indonesian/Malaysian Malay negators also largely applies to those of Modern Javanese. In this analysis, *bukan* is an external (sentential) negation marker which occurs (1) before nouns and (2) "in certain pragmatically determined contexts", including "contrastive sentences, marked narrow focus, and metalinguistic negation" (p. 137). Further-

Incidentally, Modern and Old Javanese have no negative existential negator ('to miss, lack; there is not'). These negators do occur in regional Malayo-Polynesian languages. They also occur in Javanese (and Malay) regional varieties,[6] but here they are less common and often unstable, being transitory between a negative existential and a predicative negation.[7] They also played a role in Javanese linguistic history (see Section 6.1).

In Modern Javanese, negators belonging to the above four categories and differentiated for speech register are as follows:

(2)

		Ngoko	Krama	(Madya)
1.	Predicative negator	*ora*	*botən*	(*ora*)
2.	Prohibitive marker	*aja*	*sampun*	*ampun*[8]
3.	Phasal negator ('not yet')	*duruŋ*	*dèrèŋ*	(*duruŋ*)
4.	Nominal/contrastive negator	*dudu*	*dédé, sanès*[9]	*dédé*

Their use is illustrated in the following sample sentences:

(3) Predicative negator:

Ngoko:	*Aisya*	***ora***	*tuku*	*bəras*	*duk*	*iŋ* *pasar*
Krama:	*Aisya*	***botən***	*tumbas*	*wos*	*nalika*	*iŋ* *pəkən*
	Aisya	not	buy	rice	when	LOC market

'Aisya didn't buy rice when she was at the market'

more, it is held in contrast to *tidak*, the Indonesian/Malaysian Malay "marker of internal (predicate) negation". Kroeger "provisionally assumes that *bukan* is polysemous", distinguishing an external noun negator $bukan_{N}$, and an internal verb negator $bukan_{V}$ (p. 138). His observation that Indonesian/Malaysian Malay is exceptional in lexicalising the distinction between external and internal negations (p. 137, 145) should be generalised to include Modern Javanese.

6. Cirebon Javanese (West Java) contrasts *bəli* 'not' with *laŋka* 'there is not' (Ewing 2005: 227–9). Vander Klok mentions the negative existential markers *gano?* and *géno?* in Malang Javanese (East Java) (Vander Klok 2024: 479) as well as *nana* in Tengger Javanese (Vander Klok written communication July 2025). Tom Hoogervorst draws attention to (Eastern Javanese) *ganok* (< *ga?* 'NEG' + *onok* 'EXIS') 'there is not' and *gaʈik* (< *ga?* + *aʈik* 'EXIS') 'without' (Tom Hoogervorst, written communication August 2023).

7. Indonesian/Malay *tiada* '(there is) not' is a good illustration of this transitory stage. Historically composed of a predicative negator *t(e)i 'not' (Adelaar 1992: 144) and the existential marker *ada* 'there is', it serves as both predicative and existential negator. Its contracted form *tidak* is more common and has become the default predicative negator.

8. See Section 7.3.

9. Also meaning 'other', see Section 9.

(4) Prohibitive marker:
Ngoko: ***Aja*** *ŋanti suwé*
Krama: ***Sampun*** *ŋantos daŋu*
don't wait long (time)
'Don't wait for too long'

(5) Phasal negator:
Ngoko: *Aisya* ***duruŋ*** *bali saka pasar*
Krama: *Aisya* ***dèrèŋ*** *waŋsul sakiŋ pəkən*
Aisya not.yet come.back from market
'Aisya hasn't yet come back from the market'

(6) Nominal/contrastive negator:
a. Ngoko: *iki* ***dudu*** *pipa*
Krama: *punika* ***sanès*** *pipa*
this not pipe
'This is not a pipe' (René Magritte, p.c.)
b. Ngoko: ***dudu*** *apa siŋ di-ucap-ké naŋiŋ cara*
Krama: ***sanès*** *punapa iŋkaŋ dipun-ucap-akən naŋiŋ cara*
not what REL OV-say-TV but manner
lé ŋ-ucap
ŋgènipun ŋ-ucap
NOMINALISER AV-say
'It's not what she says but the way she says it'
c. Ngoko: *jar-é sopo pit-ku ora irəŋ, siŋ koyo ŋéné*
say-3.POSS who bicycle-1 not black REL like like.this
jənəŋ-é irəŋ ***dudu*** *abu-abu*
name-3POSS black not grey
'Who says my bicycle isn't black? One like this is what you'd call black, not grey' (Susi Moeimam, p.c.)
d. Ngoko: *aku ora (/dudu) nəsu, naŋiŋ loro ati*
1 not angry but sick (in) heart
'I'm not angry, just hurt (insulted)' (Susi Moeimam, p.c.)
e. Ngoko: *kuwi dudu ñañi, naŋiŋ gəmrèwèŋ*
that not singing but humming.recital
'That's not singing, it's just some humming' (Susi Moeimam, p.c.)

Old Javanese has a very different inventory, in which a predicative negator *tan* tends to dominate in all cases where Modern Javanese applies the four negator types listed above. Moreover, Old Javanese has no speech registers (Section 3).

The Ngoko types listed above do have Old Javanese precursors, but these were less established as negators than their Modern Javanese counterparts.[10] Old Javanese *haywa* is a cognate of Modern Javanese *aja* and expresses negative imperatives, but it was clearly in competition with *tan* (Zoetmulder 1983: 169–71). Old Javanese *dudū* occasionally expresses nominal negation, but it also has various other meanings[11] and never seemed to have become an unambiguous negator. At any rate, *tan* was used more frequently for this purpose. Finally, Old Javanese does have a counterpart (and apparent cognate) *turuŋ* 'not yet' to Modern Javanese *duruŋ* (Section 8), but this word was in competition with *tapwan*, which has the same meaning and is presumably derived from *tan* (Zoetmulder 1983: 142–5).

5. Arguments for the selection of the negators under investigation

The four negators under discussion are selected on the basis of various criteria.

That they are the most common negators in Modern Javanese is easy to verify in spoken and written language, and it is also reflected in descriptions and textbooks. Other criteria are that these negators are morphologically underived, and the fact that their configuration is to some extent associated to an identifiable geographical area in Southeast Asia. In the case of phasal negators, Veselinova et al. (2024: 892–6) show that out of a corpus of 62 languages in the Malayo-Polynesian part of Southeast Asia, 18 express these negators by periphrastic constructions, 17 by a transparent or semi-transparent univerbation, 17 other languages by a single morpheme, and 10 remaining languages by using mixed possibilities. As it turns out, while the periphrastic constructions and the univerbations are not clearly representative of a certain area, almost all phasal negators expressed by a single morpheme are concentrated in the "southwest corner" of the Malayo-Polynesian language region in Southeast Asia. Here, they "form an almost perfect belt" over an area including Java, Sumatra, Bali, Lombok, Sumbawa and a few islands further east (Veselinova et al. 2024: 893; see also map on page 894).

Vander Klok (2024: 479) also lists four negators for the languages of Java but she uses a different set. In her analysis, Modern Javanese *duruŋ* is in the first place identified as a grammatical marker, the meaning of which is paradigmatically related to the phasal markers 'already' and 'still'. She therefore leaves it outside her

10. Then again, in contrast to Modern Javanese, Old Javanese does have a dedicated prosentential negator *taham* 'no' (Zoetmulder 1983: 142–5).

11. Compare Zoetmulder (1982), which has *dudū* 1. 'different, difference, distinction'; 2. 'not to mention, and furthermore; "and there were others who…"'; 3. 'wrong, not as it should be'; 4. 'not (being)…'.

treatment of negators, as do van der Auwera et al. (2024). On the other hand, she does include a category of existential negators in order to cater for Madurese and Sundanese, noting furthermore that while Modern Javanese does not have a dedicated Existential negator morpheme, some vernacular variants do. In my analysis, I do not consider existential negators as a main category, although I mention them in my discussion of predicative negators in historical hindsight, because in Javanese the latter have evolved from earlier negative existentials. On the other hand, I do include *dudu* among the Javanese negators on account of its prominence in this category in languages in the southwest corner.

6. Predicative negators

In contrast to the Modern Javanese predicative negators (Ngoko) *ora* and (Krama) *botən* which are differentiated according to register, Old Javanese only has the general negator *tan*.

6.1 The Ngoko predicative negator *ora*

The predicative negators *ora* and *botən* also function as prosentential negators (meaning 'no'). In spoken language, *ora* is often shortened to *ra?* in East Javanese, and to *ra* or *ra?* in Central Javanese.

BTS list *ora* as a reflex of PMP *wada 'be, exist, have; wealthy; not exist, not have'. The latter is the adaptation of an earlier reconstruction *vaḍa 'Vorhandensein, Nichtsein [availability; not be]' by Dempwolff (1938). The meanings attributed to PMP *vaḍa/*wada are odd because they are inherently contradictory. BTS present cognate lists for PMP *wada and for the Proto Western Malayo-Polynesian variant forms *wada-q and *waday. These are either glossed 'be, exist, have', or they are glossed as 'not exist, not have' or a closely related meaning, see Appendix 2.[12] In a note to these lists BTS comment that Dempwolff's gloss 'Vorhandensein, Nichtsein' "is difficult to improve". They also note that

> the idea of existence or possession is clearly reflected in the northern Philippines, in Iban, Malay, Mentawai, and Chamorro, while a negative sense is prominent in the central and southern Philippines, in Maloh, Old Javanese, and Modern Javanese, and in UMAT[13] of central Sulawesi. (BTS 2023: entry *wada)

12. BTS have the meaning 'wealthy' added to the gloss of *ada. However, the only base for this seems to be the occurrence of a Malay reflex *bər-ada*, which does mean 'be wealthy', in contrast to its root *ada* 'to be; to have'. I leave out 'wealthy' here, as both the form *bər-ada* and its meaning are derived.

Finally, they also observe that these opposed senses do not coexist in any attested language.

"Be" or "not be": a puzzling semantic conundrum. However, at least part of the alleged reflexes in BTS's cognate list can be discarded. In the case of the PMP *wada list, it is immediately clear that the meanings 'not exist' and 'not have' are inaccurate and should be deleted on account of the non-evidence from Embaloh ("Maloh"),[14] Modern Javanese, and Old Javanese.

Embaloh *adaʔ*[15] cannot reflect PMP *wada for two reasons. Phonologically, the consonant correspondences in *adaʔ* are irregular, because in Embaloh, PMP initial *w was as a rule maintained as *w* and PMP intervocalic *d became *-r-*. Moreover, the final *ʔ* remains unexplained (Adelaar 1994: 5, 8, 11–12, 13). From a semantic perspective, and in contrast to BTS's gloss, *adaʔ* is not a standard or predicative negator but a prohibitive marker (Adelaar 1994: 19). The Embaloh default predicative negator, *naʔan*, is formally very different. Typologically, standard (/predicative) negators and prohibitive negators are notably distinct (Sadock & Zwicky 1985: 175; Aikhenvald 2010: 165).

As will become clear below, *ora* 'not; no' does not derive from PMP *wada but from a reduced form of Old Javanese *tan* 'general negator' + *wwara* 'existential marker' (*wwara* by itself reflects PMP *wada, see below). As documented in various Javanese literary texts, throughout the history of this language, the existential negator *tan wwara* underwent "univerbation" to *tan-wwara* and became gradually reduced to *tanora, nora*, and *ora*. All of these forms are attested in Zoetmulder (1982). The final stage form *ora* has become a predicative negator: it no longer expresses existential negation except in combination with the current Modern Javanese existential marker *ana*. One of the factors contributing to the gradual loss of *tan* must have been its tendency to undergo univerbation with following verbs, causing the addition of an extra syllable. This would trigger syllable reduction, setting into motion the neutralisation of the antepenultimate vowel to *ə* and its subsequent reduction to Ø, which are both regular sound changes in Javanese (Nothofer 1975: 71) and Malay (Adelaar 1992: 49–54).

The negator *ora* 'not' also occurs in Zoetmulder's (1982) Old Javanese dictionary, which probably explains why it is listed in BTS as a reflex of PMP *wada for

---

13. This word in capitalised spelling is somewhat puzzling. It must refer to Uma, a language spoken in Central Sulawesi which has *araʔ* 'there is none'. The latter word is listed in BTS, see Appendix 2.

14. "Maloh" often appears in the linguistic literature, but this is an Iban exonym. The people who are designated with it prefer to call themselves and their language "Ambalo" or "Embaloh" (Adelaar 1994).

15. The word has a final glottal stop (Adelaar 1995: 380).

both Old and Modern Javanese. However, it is by no means a default predicative negator in Old Javanese, which has *tan* instead (Zoetmulder 1983: 143; see also van der Molen 2015). In Zoetmulder (1982), which is a philological dictionary, it only occurs in sample sentences from Middle Javanese texts.[16]

The way in which Old Javanese *tan* was replaced by Modern Javanese *ora* is reminiscent of the Jespersen Cycle (Dahl 1979), which is often demonstrated on the basis of the development from Old French *ne* to Modern French *(ne) … pas*.

In Old French, the noun *pas* 'step, pass' was initially added as an emphatic marker to *ne* forming *ne … pas* 'not [even] a step' (*je ne sais* 'I don't know' > *je ne sais pas* 'I don't know at all'). Later on, *ne … pas* became the standard default negator in Standard French (> *je ne sais pas* 'I don't know'), whereas in modern Colloquial French, there is a strong tendency to drop *ne* and use only *pas* as a default standard negator (> *je sais pas* 'I don't know'). See van der Auwera et al. (2022), who furthermore point out that in standard French this development actually does not represent a meaning shift of the negation from *ne* to *pas* but rather one from *ne* to the discontinuous co-occurrence of *ne* and *pas* (that is, [*ne … pas*]) within the same verb phrase. This is because in other contexts, *pas* has maintained its original meaning 'step, pass'.[17]

In Javanese linguistic history, the development from *tan* to *ora* went via an intermediate *tan wwara* (Kern 1883: 22–23). Old Javanese *wwara* is an existential marker and a regular reflex of PMP *wada (Nothofer 1975: 84). It was replaced by *ana* in Modern Javanese.

The development from *tan* to *ora* is in many ways similar to that of French *pas*, but it differs from it in that it specifically involves an existential verb. It is therefore more in line with Croft's (1991) Negative Existential Cycle, which has been the topic of various other publications including Veselinova (2014, 2016), Veselinova and Hamari (2022), van der Auwera, Krasnoukhova and Vossen (2022), Vossen and van der Auwera (2014). Veselinova and Hamari (2022: 2) characterise it as a possible pathway for the development of standard negation markers from negative existentials. In his discussion of the Negative Existential Cycle, Croft (1991: 6–11) distinguishes three main language types (along with three intermediate types). Following Veselinova and Hamari's (2022: 8) characterisation of these, Type-A languages use the same negator in both predicative and existential clauses (whereby it is used in combination with the existential marker in existential clauses); Type-B

16. These texts include *Kiduŋ Sunda, Sorandaka, Korawāśrama, Waseŋ Sari*. I am grateful to Stuart Robson for pointing this out to me.

17. Original meanings are also maintained in some other French lexemes which became emphasisers in combination with *ne* (e.g. *personne* 'nobody; also: person', *point* 'not at all; also: point', *jamais* 'never; also: ever').

languages have a dedicated existential negator; in Type-C languages, the negative existential is identical with the standard negator but they appear in different constructions.[18]

Croft distinguishes three main Cycle stages for Type-A languages:

Stage One:	there is a general negator for both standard and existential negation
Stage Two:	it becomes a default marker for standard negation, whereas there is a new marker for existential negation
Stage Three:	the existential negation marker becomes a standard negator; in combination with a (new) existential marker, it is also used for existential negation.

These stages can also be observed in the development from Old Javanese *tan* to Modern Javanese *ora*, which conforms to a Type-A structure:

In *Stage One, tan* is a default negator used for both predicative and existential negation. There is no dedicated single morphemic negative existential marker, negative existence being expressed by *tan* + the existential marker *wwara*: Old Javanese *tan* 'no, not' + *wwara* 'be there (/have)' > *tan wwara* 'not be there (/not have)'.

In *Stage Two*, the collocation *tan wwara* becomes a marker of predicative negation. It is in the process of becoming fused into one word and further reduced to *tanora, nora* and *ora*, intermediate forms which still occur in late Old Javanese and Middle Javanese texts: Old Javanese *tan wwara* 'not be there (/not have)' > late Old Javanese/Middle Javanese) > *tanora* > *nora* > *ora* 'not be there (not have); not'.

In *Stage Three*, the eventual form *ora* has become the predicative negator in Modern Javanese, which by now has lost all phonological trace of the original negator *tan. Ora* is not an existential negator by itself: in order to express existential negation, it needs to be followed by the (current) existential marker *ana*, as demonstrated in the following sample sentence:

(7) Modern Javanese (Elisabeth Riharti p.c.)
dina iku kok ora ana bocah iŋ səkolah
day that PART [+wonder] NEG EXIS child LOC school
'I was wondering why there were no children in school that day'

Without existential marker *ana* the sentence would be incorrect:

+dina iku kok ora bocah iŋ səkolah

Two observations should be made with regard to the history of *ora*. First, in Javanese, the Negative Existential Cycle has gone full circle. While the Cycle distinguishes three stages, van der Auwera et al. (2022: 613) and Veselinova and

18. This characterization is somewhat more explicit than the one offered by Croft (1991).

Hamari (2022: 3) note that a completion of the last stage is rarely observed. However, in the Javanese history, it appears that all three stages have been fulfilled. In an initial stage, Old Javanese *tan* was a predicative negator; it only expressed existential negation in combination with the default existential marker *wwara*. In a second stage, *tan wwara* was a predicative negator as well as an existential negator (undergoing "univerbation" and becoming reduced to *tanora* > *nora* > *ora*. And in a final stage, *ora* has become a predicative negator; it only expresses existential negation in combination with the current existential marker *ana*.

Second, Veselina and Hamari (2022: 3) furthermore note that the Cycle is rarely fulfilled "within the time span for reasonable construction". However, Veselinova's (2014) time frame for the completion of the Negative Existential Cycle calls for an adjustment. Primarily using evidence from Polynesian historical linguistics, she estimates "the observable time [for the completion of the Negative Existential Cycle] to be about 2,000 years, the time specified for the breakup of Polynesian varieties into separate languages." (Veselinova (2014: Section 5).[19] However, at least in the case of Javanese predicative negators, which have gone full circle from *tan* to *ora*, the time depth is less than a millennium. The earliest Old Javanese text is an inscription from 804 which was found in Sukabumi (Kediri Regency, East Java; Zoetmulder 1974: 3), while Modern Javanese as we know it today is basically the literary standard which emerged at the courts of the sultanates of both Solo and Yogyakarta in the 18th–19th century (Ogloblin 2005: 591). This suggests less than half of the time implied by Veselinova's estimated *ante quem* period. The time estimated for the *tan*/*ora* cycle admittedly remains somewhat elusive, given the uncertain relationship between Old Javanese and Modern Javanese. For instance, the year 804 may not be significant as a starting point of the cycle in a dialect which in all likelihood was not directly ancestral to Modern Javanese (Section 2). Nevertheless, given the available evidence for the history of Old Javanese and Modern Javanese in general, it is safe to assume that the *tan*/*ora* cycle took much less than two millennia to come full circle.

19. This evidence is backed up by preliminary observations made by Veselinova on the basis of Uralic data.

6.2 The Krama predicative negator *botən*

The Krama counterpart of *ora* is *botən*, which must have originated from *tan* + *wontən*.[20] In turn, *wontən* is paradigmatically the Krama counterpart of Modern Javanese *ana*, and is historically derived from (Old Javanese) *wwara*.

This etymology may not be transparent, but neither is it irregular. It is supported by the intermediate Old Javanese forms *tan bontən, tambontən, mbontən* and *bontən* 'not to be; not'. It is also supported by the circumstance that the Krama counterpart of *ana* is *wontən*, which in turn is a later form of *wwantən* (still extant in Old Javanese). Although phonologically odd, the formal relationship between *wwara* and *wwantən* follows one of the common Kramafication processes causing the last syllable of a Ngoko root to be replaced by *-ntən* (or, in the case of *botən*, by *-tən*) in the corresponding Krama form (Section 3).

(8)	Ngoko	Krama
	ana (which replaced *wwara*) 'to be'	*wontən* (< *wwantən*) 'to be'
	ora (< *tan wwara*) 'no, not'	*botən* (also *tambontən, mbontən, bontən*) 'not to be; not'

6.3 Some remaining questions

The etymology of *botən* is likely, but it leaves some phonological changes unexplained. For instance, why was *ww* lost in the change from *tan wwara* to *tanora* (> *ora*), whereas it underwent fortition to *b* in the change from (*tan) *wwantən* to *botən*? While there is no comprehensive answer to this question, the least that can be said is that in the history of Old Javanese root-initial *w* or *ww*, both lenition and fortition appear more generally when *w*/*ww* is directly preceded by another consonant:

Lenition of *w-* or *ww-* to *o* (sometimes through contraction with following *a*), or loss of *w-*:

(9) *tan wwara* > *tanora* > Modern Javanese *ora* 'no, not'
wawa 'carrying along (root)' (Zoetmulder 1982) > *a-mawa* 'to bring along' (Nothofer 1975: 125)
saŋ (honorific article) + *hyaŋ* 'god, divinity' + *wai* 'sun; day' > *srəŋ-eŋ-we > *srəŋéŋé* 'sun' (Gericke et al. 1901)
sor 'low position' + *wai* 'sun; day' > *sor-we > *soré* 'evening' (Adelaar, to appear)

20. I am grateful to Arlo Griffiths for having drawn my attention to this etymology. Compare Kern (1883: 22–23) for sandhi cases involving *tan*.

Fortition of *w-* or *ww-* to *b*:

(10) *tan wwantən, tan wontən* (Krama form of *wwara* 'exist') 'not exist' > *tambontən* > *botən* 'no, not'
tan + *wəruh* 'know' > *tambuh* 'not know' (with loss of penultimate syllable)
tan + *wis* 'already' > *tambis* 'almost' (Zoetmulder 1983: 23)
ŋ-ombé 'to drink' possibly from: *inum* 'drinking' + *wwai* 'water' > *<in>umwé > (+ re-analysis) > *umbé > *ŋ-ombé* 'to drink'[21]

Another irregularity for which I still owe an explanation is the lack of an (expected) pre-consonantal *n* in *botən*: why does *wwantən* have one, but *botən* does not?

Tom Hoogervorst (p.c.) points out the Old Javanese variant *wwaya* 'to be present' and the (now obsolete) Modern Javanese form *boya* 'not' as a possible case of an existential marker becoming a general negator.[22] Formally, the way they match is no less straightforward than the way *wwara* does in relation to forms like *ora* (and *botən*), if we assume that *boya* derives from an earlier combination of *tan* and *wwaya* which subsequently underwent deletion of *tan* after it had become obsolete and had caused fortition to the following consonant. The development would have been as follows: *tan* + *wwaya* > *tan-w(w)oya > *tam-boya > *boya*. The histories of *wwaya* and *boya* remain obscure: although *wwaya* is listed in Zoetmulder (1982) and *boya* in Pigeaud (1938), the scholarly literature does not throw further light on their origins.

7. Prohibitive markers

As already indicated, both Old Javanese and Modern Javanese have prohibitive markers, which in Modern Javanese are differentiated for high and low register. In what follows I first discuss the Ngoko form *aja* together with Old Javanese *haywa*, as they have a common source (Section 7.1). I then evaluate an alternative explanation of the origin of *aja* by BTS (Section 7.2), before directing my attention to the Krama form *sampun* (Section 7.3).

21. Author's own etymology.

22. The forms are also mentioned in Kern (1883: 22–23).

7.1 Modern Javanese Ngoko *aja* and Old Javanese *haywa*

I argue that the prohibitive markers *aja* and *haywa* are related to Old Javanese *hayu*/*ayu*.[23] More precisely, they are phonologically derived from a subjunctive form *hayu-a of this root. The root itself is used verbally as well as nominally and conveys various meanings related to goodness, well-being, rightness, beauty, agreement and harmony. It has a corresponding form *ayu* 'beautiful, pretty' in Modern Javanese. I furthermore claim that Old Javanese *hayu*/*ayu* and Modern Javanese *ayu* are also historically related to the Modern Javanese hortative marker *ayo* 'c'mon, let's!' (see further below).

In its capacity of a stative verb, the root *hayu* (/*ayu*) covers the meanings '(be) good, well, right, beautiful, happy'. Phonologically, *haywa* and *aja* are contracted forms of *hayu-a. This is most apparent in *haywa*, whereas *aja* is a more recent form and has evolved further away from it. (The latter is also part of spoken language, whereas Old Javanese *haywa* belongs to a predominantly literary language). The evolution of *aja* must have been as follows: *hayu-a > *haywa* > *(h)aya > *aja*. Modern Javanese word structure disfavours heterorganic consonant clusters, and a cluster of semivowels like *yw* is particularly awkward, hence its reduction and fortition to *j*.

The fortition of *y* in Modern Javanese is a common sound change, although it is not exceptionless. Other examples (which are more straightforward than *haywa* > *aja*) are:

(11)		Old Javanese	Modern Javanese
	a.	*layū, a-layū, ma-layū* 'to run, flee' (also: *laju* 'going on, further, and forthwith'	*laju* 1. 'straight (on), forward, continuous; 2. and then, after that'; *ŋlajo, ŋlaju* 'to commute from (home)'; also *layu* 'to run'
	b.	*yasa* 'worth, honour, renown'	*jɔsɔ* (regional) '(acquired) merit'; *yɔsɔ* 'building' (ultimately from Sanskrit, compare *yaśa* 'renown, honour, meritorious deed', also *jaśas* 'public works' Gonda [1973: 419])

23. In Zoetmulder (1982), *hayu* and *ayu* are given separate entries with different meanings, but I treat them as formal and semantic variants of one and the same root *hayu*. Zoetmulder (1982) has the entries *hayu* 1. and *ayu* 1., which cross-refer to one another. *hayu* 1. is glossed '(the) good, goodness, rightness, virtuousness, good condition, well-being, welfare, happiness, felicity; beauty, loveliness'; the derivations of *ayu* 1. (which does not occur as a root by itself) are given meanings conveying 'being in harmony', 'being agreed upon', 'matching', and also 'partnering' and 'competing'. Zoetmulder treated *hayu* 1. and *ayu* 1. as separate roots, as he did with the prohibitive marker *haywa* and its variant form *aywa* as well as with *hayo* and *ayo* (hortative markers). However, I assume that all these words are phonological variant forms and/or are morphologically related, whereas their meanings are clearly polysemous. I assume that they derive from the same basic root *hayu which could be used as both a noun and an adjective ('good').

c.	*Yawa* 'Java; Javanese'	*Jɔwɔ* 'id.'
d.	*yawi* 'outside'	*jawi* '1. 'outside; 2. Java(nese)'
e.	*yogya* 'proper, suitable; able'	*Yogyakarta, Jogyakarta* 'name of city in Java'

As I discuss further below, the fortition of *y* did not happen in *ayo*. Semantically, it is necessary to explain how a word meaning 'goodness, being good, rightness, beauty, well-being, agreement and harmony' ends up in directive clauses, becoming a prohibitive marker no less. To be more precise, it is necessary to explain how the subjunctive marker *-a* is used in Old Javanese, how subjunctives can be used in directive constructions and end up in prohibitive ones, and how *hayu-a* with its originally positive meaning could have become a negator.

To begin with, in Javanese, *-a* can derive dynamic verbal clauses with a directive meaning from roots that are not verbal: in this way, a deictic marker *mréné* 'hither' may become *mréné-a!* which has a directive meaning '(Please) come this way!'. Likewise, *-a* suffixed to *hayu* in its capacity as an attribute yields *hayu-a* 'it would be good/right/beautiful/agreeable if…'.

This brings us to the use of the subjunctive to form prohibitives. In all likelihood, the derivation *hayu-a* would have acquired a desiderative meaning 'it would be better if', or 'you'd better', 'why don't you…' and be used in polite orders: 'please (be so kind as to)…', and this polite order would have become a prohibitive, most likely (but not necessarily) in combination with a negator (see below): *hayu-a 'it would be good (if)…' (a wish) > 'please do…' (a polite order), which, in combination with a negator, becomes a prohibitive. The use of the subjunctive in prohibitive utterances is typologically common. It is a less direct and generally more polite form of giving orders. As a strategy for not "spelling out" things, it is motivated by a need to avoid the risk of losing face, or of causing the loss of it, which is inherent in the use of directives, and more particularly in negative ones (Aikhenvald 2010: 351–364, 398–410; Van Olmen 2010: 484–502; Devos & Van Olmen 2013: 25–30). A clear instance in European languages is Spanish, which has a dedicated imperative but uses the standard negator *no* in combination with the subjunctive mood in prohibitive phrases, see the contrast between imperative *deja-me!* 'leave me!' and prohibitive *no me dejes!* 'don't leave me!' (see also Aikhenvald 2010: 193, 398).[24] Devos and Van Olmen show that the subjunctive is also frequently used in prohibitive phrases in Bantu languages. In reference to the evolution of Javanese *aja/haywa*, Zulu is particularly interesting as it has a prohibitive marker *kahle* which also occurs as an adverb meaning 'well, in a good

24. In these phrases *deja* is the imperative of the verb *dejar* 'to leave, abandon', and *dejes* is the 2nd person singular subjunctive; *me* is a 2nd person singular object pronoun; *no* is a standard negator.

way' or 'carefully', and it is also used as a directive meaning 'wait!' or 'be careful!' They point out that *kahle* has apparently gone from a positive to a negative sense, "possibly via the intermediate step of a warning" (Devos & Van Olmen 2013: 36).

There are several speculative explanations for the development from *hayu-a 'it would be good if…' to *haywa* and *aja* 'don't!':

a. Analogous to the way *tan* + *wwara* 'not be' ended up as *ora*, *hayu-a was initially also preceded by *tan* (> *tan* *hayu-a) which was lost after it had become obsolete as a result of antepenultimate vowel reduction and a shift of its meaning to the following *hayu-a. Compare:

   (12) *tan wwara* > *tanwwara* > *tanora* > *nora* > *ora*
   not be > not-be > not-be, not(?) > not-be, not(?) > not

   (13) tan *hayu-a > (tan) *haywa > *haywa* > *aja*
   not + would be good > (Ø) would not be good > better not, don't > don't

b. Alternatively, it is possible that *haywa* was never negated and went directly from 'it would be good if…' to 'don't!', in analogy to what may have happened with Zulu *kahle*.
c. As suggested by the reviewer of an earlier draft of this chapter, *haywa*/*aja* may be derived from *hayua… tan… 'it would be better if not…' rather than from *tan hayua 'it is not the case that it would be better if…', arguing that one would expect the former construction to have a more convincing syntactic-semantic scope. However, some factors mitigate against this analysis. As already argued, positing *tan* in front of *hayu-a provides a more plausible scenario as it would bring the evolution of *haywa/aja* in line with the evolution from *tan wwara* to *ora*. Furthermore, the pressure of antepenultimate vowel neutralisation is rather prevalent in Javanese word structure, providing a plausible explanation for the loss of *tan* if it historically had preceded *hayu-a and then had merged with it through univerbation. However, neither this scenario nor antepenultimate vowel neutralisation would be capable of explaining *tan* in these ways if it had followed *hayu-a. Finally, there are in fact Old Javanese constructions combining *haywa* with a following *tan*, but these do not express negation. Zoetmulder (1983: 171) characterised such combinations as "a firm confirmation of an invitation or order", as in the clause *haywa tan wiwéka* 'use your brain' (*haywa* 'don't'; *tan* 'no(t)'; *wiwéka* 'insightful, with discrimination').

After *hayu-a had changed into the prohibitive markers *haywa* and *aja*, they were no longer associated with the root *(h)ayu*, as is reflected in Zoetmulder (1982), Pigeaud (1938) and Robson and Wibisono (2002), who all treat *hawya*/*aja* and

(h)ayu as unrelated roots. This disconnect is a precondition for re-analysis to take place (Harris & Campbell 1995: 65–71).

At this point it is worth recalling that *aja*, being a later form than *haywa*, probably did not evolve in a direct line from the latter (Section 2). Some morphosyntactic differences between Old and Modern Javanese clearly suggest that Modern Javanese did not directly evolve from Old Javanese but from a dialect which existed alongside it (Adelaar 2011). Ogloblin (2005:591), referring to Pigeaud (1967) outlines how Modern Javanese came up as an independent literary style in 16th century East Java, from where its centre shifted early onwards to Java's north coast (a centre of Islamic writing) and later on again to the Central Javanese hinterland. In Central Java, a variety of it first used at the court of Solo became a literary standard at the courts of the sultanates of both Solo and Yogyakarta in the 18th–19th century. It also became the basis of standard Modern Javanese.

The point is worth making because the notion of dialect variation in the past may explain the existence of yet another ("doublet") reflex of *hayu-a, namely Old Javanese *hayo* or *ayo* and its Modern Javanese counterpart *ayo*. In Old Javanese, *hayo*/*ayo* is a variant form of the prohibitive marker *haywa*; however, in Modern Javanese, *ayo* (colloquially also *ayuh* or *ayuk*), is a hortative marker ('c'mon! let's!' or, in an invitation to do something, 'please!'). Phonologically, the resulting forms have often contracted the root-final *u with *a to *o*: *hayu-a > *hayo, ayo*. Compare phrases (14)–(18):

(14) Old Javanese (Zoetmulder 1982)
ayo masku trəsneŋ tuwuh!
ayo mas-ku trsna iŋ tuwuh
PROH gold-my desire PREP growing; life
'don't cling on what's alive, my precious!'

(15) Old Javanese (Zoetmulder 1982)
ayo göŋ krodha
PROH strong(ness) anger
'don't be too angry!'

(16) Old Javanese (Zoetmulder 1982)
hayo ta wineh masaŋyoga
hayo ta <in>weh ma-saŋyoga
don't EMPH <UV>give INTR-have.intercourse
'Don't let [so-and-so] have sexual intercourse!'

(17) Modern Javanese (Elisabeth Riharti, p.c.)
Ayo ndaŋ ñambut gawé!
HORTATIVE hurry do work
'Come on, let's do some work!'

(18) Modern Javanese (Elisabeth Riharti, p.c.)
Ayo maŋan saiki!
HORTATIVE eat now
'C'mon let's eat!'

Having maintained the semivowel *y, *hayo* and *ayo* seem formally closer to *haywa* than to *aja*, but it still remains unclear whether they are more akin to Old Javanese or to Modern Javanese, (or perhaps, whether they originate from yet another hitherto unidentified dialect). The semantic relationship between *hayu-a 'it would be good if…' and a positive directive like *ayo* 'c'mon! let's', is much more straightforward than the semantic relationship between *hayu-a and the prohibitives *haywa* and *aja*, and there is good reason to assume that these forms are related. Their formal differences may be the result of a combination of regular sound change, irregular sound change due to frequency (ISCDF, Section 1), and dialect variety. They must be the reason why Old Javanese and Modern Javanese together have a large number of different phonological realizations of *hayu*, including *hayu, ayu, haywa, aja, hayo, ayo*, carrying meanings varying from 'being good, beautiful, valid, proper', to being hortative or a prohibitive marker. Compare the following developments:

(19) *hayu* 'beautiful, virtuous' > *(h)ayu* 'beautiful, virtuous'
hayu > *(h)ayu-a* 'it would be beautiful/good if…' > *(h)ayo* 'c'mon, let's! (hortative)'
hayu > *(tan) (h)ayu-a* 'it would not be good if …' > *(h)ayo, haywa, aja* [ɔjɔ] 'don't'

Mańczak's (1980) ISCDF theory holds that frequently occurring "linguistic elements" (words, word groups and morphemes) in a language are subject to irregular sound change, depending on their length and frequency of occurrence. Length and frequency of occurrence are in a precarious balance, but since either of them may change (e.g. length may reduce as a result of regular sound change), this balance is often disturbed to the effect that if a linguistic element becomes too short in relation to its frequency, it will be replaced by a longer one, and if it becomes too long, it will be shortened (Mańczak 1980: 63). According to Mańczak, ISCDF happens in every language and affects about one third of its linguistic elements. He also argues that, if a linguistic element occurs both in a regular and an irregular form, ISCDF applies more often to the irregular form than to the regular one (p. 65). The occurrence of competing forms one of which is phonologically regular while the other one is not (or several other ones are not), must be one of the factors (together with dialect variation) contributing to the various current reflexes of *hayu a.

7.2 BTS's explanation of the origin of *aja*

BTS derive Modern Javanese *aja* from a PMP prohibitive marker *ədaq, which presumably has a Proto Western Malayo-Polynesian doublet form *ənda, a (standard?) negator glossed 'no, not'. (They do not deal with Old Javanese *haywa*). Their evidence for *ədaq and *ənda include the following reflexes:

(20) PMP *ədaq 'vetative,[25] don't' >
Maranao *daʔ* 'none, nil, lose, not'
Tiruray *ʔendaʔ* 'none, not any, not'
Embaloh *adaʔ* 'don't'
Brunei Malay, Sarawak Malay *endah* 'no, not'
Sasak *endaʔ* 'don't'
Taeʔ *da(ʔ)* 'negative imperative: don't'
Javanese *aja* 'don't; may it not'
Bugis *ennaʔ* 'no, not'
Kambera *nda* 'particle of negation'

(21) Proto Western Malayo-Polynesian *ənda 'no, not' >
Melanau *nda* 'no, not'
Iban *ndaʔ* 'not'
Balinese *eda* 'particle marking a negative imperative; do not'

These etyma are problematic and do not provide a satisfactory alternative explanation for the origin of *aja*. Without venturing too far into their status, a phonological comparison of the twelve reflexes adduced by BTS in their support leaves some irregular sound correspondences that are difficult to account for. To address the most important ones, initial *ə does not agree with Javanese *a*, and *d should have yielded Javanese *r* (at least in initial and intervocalic position). Furthermore, as pointed out in Section 6.1, Embaloh *d* and Javanese *j* do not reflect PMP *d.[26]

The twelve reflexes also represent a conflation of prohibitive markers and other (standard?) negators, which is semantically unwarranted. Moreover, only four of them are labelled as prohibitive markers, and of those four, the Embaloh

25. The term "vetative" (or "vetitive") refers to events which the addressee is warned against happening. In the current context it is not entirely accurate. As it is used in juxtapunction with 'don't', it is almost certainly meant to refer to the notion "prohibitive", with which it is often confused.

26. The fact that in the original text, *aja* (as well as Buginese *ennaq*) occur in a note section suggests that Blust himself — the initial author — also had some reservations about the derivation of these words from *ədaq.

and Modern Javanese reflexes are clearly not related to the other reflexes through regular common inheritance (Section 6.1).

Finally, the reflexes in Malayic languages (Brunei Malay, Sarawak Malay and Iban) are the result of the reduction of earlier compound forms involving the Proto Malayic negator *n[i] followed by the existential marker *ada (van der Auwera et al. 2024: 877).

7.3 The Modern Javanese Krama prohibitive marker *sampun*

The Krama word *sampun* has two main functions. It is basically the counterpart of the Ngoko phasal polarity marker *wis* 'already'. In this role it applies to 'an (expected) situation that has come to exist' (Vander Klok 2024: 485; Veselinova et al. 2024: 889), as illustrated in the following sentences:

(22) Modern Javanese (Ngoko) (Elisabeth Riharti p.c.)
Kanca-kancaku wis maca warta
kanca-kanca-ku N-waca
friend-RDP-my already AV-read news
'My friends already read the news'

(23) Modern Javanese (Krama) (Elisabeth Riharti p.c.)
Paduka sampun maos pawarta
N-waos
majesty already AV-read news
'Her majesty already read the news'

However, it also has a derived meaning, which functions as the Krama counterpart of the Ngoko prohibitive marker *aja*. Compare:

(24) Modern Javanese (Ngoko) (Elisabeth Riharti p.c.)
Aja ŋomoŋ saduruŋé dikon!
aja N-omoŋ sa-duruŋ-é di-kon
PROH AV-talk as.long.as-not.yet-of.it UV-ask.to
'don't talk until asked to!'

(25) Modern Javanese (Krama) (Elisabeth Riharti p.c.)
Sampun makatən!
PROH like.that
'Don't be like that!'

(26) Modern Javanese (Krama) (Elisabeth Riharti p.c.)

Sampun	*kəsupèn*	*ɖahar,*	*ŋgih!*
	kə-supé-(ə)n		
PROH	INVOL-forget-INVOL	eat	OK
Dipun	*təlasakən*	*ɖaharanipun*	
	təlas-akən	*ɖaharan-ipun*	
UV	finish-APPL	meal-ANAPH	

'Don't forget to eat, OK? Better finish your meal…'

In the sentences containing *sampun*, its basic meaning is clearly 'already, finished', which it shares with its Ngoko counterpart *wis*. Starting from there, it acquired a cessative meaning 'it's finished! finish/stop doing…!' before it became a general prohibitive marker 'don't…!' ('already, finished' > cessation > prohibitive). Compare sentence (27), in which *sampun* still has a cessative meaning:

(27) Modern Javanese (Krama) (Elisabeth Riharti p.c.)
Sampun!
PROH (CESSATIVE)
'That's enough!'; 'Stop it!'

Similar expressions are French *C'est fini!*, German *Schluss*! 'It is finished!', 'It's over!', or the Italian, Spanish, and (via borrowing) English *Basta*! 'Enough!'. In many of these examples the cessative marker occurs by itself. However, it does not have to occur in isolation, as examples like 'Stop talking' (versus 'Don't talk') clearly show.

Aikhenvald considers the cessative as a subcategory of the prohibitive, which is used to bring an action or situation to an end (to make it "stop"). She considers cessative constructions to be among the major pathways for developing prohibitives (2010: 358–359).[27]

Incidentally, one wonders about the accuracy of a parity between *aja* and *sampun*, considering that the use of Krama register implies respect, which to some extent precludes the bluntness of a prohibitive. In practice, prohibitives in Krama register must be much milder and circumspect (and possibly more directed towards the well-being of the addressee) than prohibitives in Ngoko register, which can be blunt.

Finally, note the Madya form *ampun*. Like Krama *sampun*, it also assumes the roles of prohibitive marker as well as phasal polarity marker meaning 'already, finished' (Robson & Wibisono 2002).

---

27. Devos and Van Olmen (2013: 34) reach a similar conclusion with regard to Bantu languages, in which cessative auxiliary verbs are the primary source for the development of prohibitives, followed by verbs expressing 'to know', 'not want', 'to come', 'lack', and 'fail'.

8. Phasal negators

Modern Javanese has the phasal negators *duruŋ* and *dèrèŋ*/*sanès* which are respectively Ngoko and Krama variants. *Duruŋ* and *dèrèŋ* show a similar (although not identical) vowel contrast as do the low and high register forms *dudu* and *dédé* discussed above. Robson and Wibisono (2002) also give a Modern Javanese regional variant form *uruŋ*.

Old Javanese has *duruŋ* 'not yet', which occurred 'in *kid[ungs]* and later prose works' (Zoetmulder 1982). *Kidungs* refer to songs as well as to a literary genre which came into existence in the late 13th century CE (Willem van der Molen p.c.). The negator *duruŋ* furthermore has the variant forms *dèrèŋ* and *turuŋ* (Zoetmulder 1982).

The origin of *duruŋ* is unclear, although it may be related to Modern Javanese *wuruŋ* 'to fail to happen' and to its Old Javanese cognates *wuruŋ* and *uruŋ*, both meaning 'failing, failing to become or take place, being unsuccessful, remaining unfinished, frustrated'. Semantically, the notions of 'not yet' and 'fail to happen' are contingent, one ('not yet') referring to an event that is still to happen, and the other ('failing to take place'), to an event that is no longer expected to do so. Formally, the loss of *w* in *uruŋ* is not unexpected, as this consonant is articulatorily weak and has a tendency to disappear in initial position, especially where it is preceding a back vowel. This is seen in pairs like *wutah* and *utah* 'to spill out'; *wutuh* and *utuh* 'whole, intact, unbroken'; *wuda* and *uda* 'naked'; *wuruh* and *uruh* 'scum, foam'. It is also seen in historical perspective in the developments from *wwara* to *ora* (with colouring of the following vowel) and from *wontən* to *ontən* (Section 6.1). The *d-: t-* contrast in *duruŋ* and *turuŋ* is odd, although there are other instances that are similar.[28] In fact, it is relatively minimal as it is not one of voice but involves tense versus lax pronunciation. (In other words, Modern Javanese *d-* is voiceless, see further Hayward 1993, 1995).[29]

28. Compare regional Javanese *pok* versus *(m)bok*, a 2nd person clitic (Vander Klok written communication 15th July 2025).

29. For obvious reasons, a detailed account of the phonetics of Old Javanese *t-* remains beyond our reach.

9. Nominal/contrastive negators

Javanese has three nominal/contrastive negators which (to some extent) are differentiated according to register:

- a Ngoko form *dudu* '1. not, other than, different; 2. wrong, not as it should be; 3. no (prosentential negator);
- a Madya and Krama form *dédé* 'no, not' [...]
- a Krama form *sanès* '1. different, other'; '2. no, not, other than, not the same thing as' (it is the Krama equivalent of *dudu*); '3. Other' (it is the Krama equivalent of *liya*) (Robson & Wibisono 2002).

Corresponding to these several forms, Old Javanese only has *dudū* 1. 'different, difference, distinction'; 2. 'not to mention, and furthermore; "and there were others who..."'; 3. 'wrong, not as it should be'; 4. 'not (being)...' (Zoetmulder 1982). It formally differs from its Modern Javanese cognate *dudu* in having a long final vowel. Semantically both cognates are very similar, although *dudu* is more decidedly a negation, whereas the meaning of *dudū* appears to lean somewhat more to that of 'other' and 'different'. Moreover, as a negator of nouns, the role of *dudū* was overshadowed by that of *tan* (Zoetmulder 1983: 142–3). The polysemic range these cognates cover is very similar to that of Proto Western Malayo-Polynesian *bukən 'negator of nominals' and its doublet *bəkən 'negator of nominals; other, different', which are listed in BTS. Indonesian/Malay has *bukan* 'not (of alternatives), no (+ noun)', the origin of which is demonstrably from an earlier (Proto Malayic) *bukən 'other' (Adelaar 1992: 76, 143). The direction of the semantic shift that took place in some reflexes of *bukən/*bəkən/*bukan* and *dudu/dudū* most likely went from 'other, different' to a nominal/contrastive negator.[30] The same semantic shift must have happened in Sundanese *laen* 'not this one (of alternatives)' (Eringa 1984) which is a cognate of Indonesian/Malay *lain* 'other, different' (Stevens & Schmidgall-Tellings 2004).

As for the high register forms, *dudu* and *dédé* have decidedly more negator functions than *sanès*, which is more polysemous. Formally, *dédé* is directly derived from *dudu*: the change of vowels conforms to one of the more often occurring processes of Krama word formation, favouring front vowels over back vowels, which is also observed in the set *duruŋ/dèrèŋ* (Section 4). I have not been able to trace the origin of *sanès*.

30. Compare also the Indonesian/Malay expression *omoŋ yaŋ bukan-bukan* [talk which not-not] 'nonsense, claptrap', in which *bukan-bukan* 'making no sense' is easier to explain if *bukan* originally meant 'other'.

10. Concluding remarks

In this paper I have tried to trace the formal and semantic origins of the main negators in Old Javanese and in the Ngoko and Krama speech registers of Modern Javanese.

Based on data from Old Javanese literary texts, I was able to demonstrate that the Modern Javanese low predicative negator *ora* and its high counterpart *botən* derive from *tan wwara*, a collocation consisting of the Old Javanese general negator *tan* which was directly followed by its existential marker *wwara*. The latter is a regular reflex of the PMP existential marker *wada. The negator *tan* was gradually lost in the process, as observed in the texts.

This loss was not clear to previous researchers, who had a difficult time understanding the presence of contrasting meanings 'to be, exist' in some languages and 'not' or 'not (to be)' in other ones, and felt compelled to reconstruct both meanings for PMP *wada. However, a closer look at the philological and phonological evidence establishes that the somewhat esoteric semantic reconstruction they made as a result is not warranted if based on the evidence of Javanese and various other languages.

In order to show how *tan wwara* 'there is not' became *ora* 'not', I used the Negative Existential Cycle model developed by Croft (1991), Veselinova (2014, 2016) and Veselinova and Hamari (2022). Two things became clear in the process of doing so. While only few languages seem to have fully undergone the Cycle, Modern Javanese appears to have gone all the way from Old Javanese *tan wwara* 'not be' to *tanwwara, *tanora, nora* and *ora* 'not; not be' and on to Modern Javanese *ora* 'not'). Furthermore, whereas Veselinova (2014) on the basis of Polynesian historical linguistic data estimates the time span for the Negative Existential Cycle to take place at two millennia, philological evidence from Javanese suggests that this language must have gone full circle in closer to half that time.

The prohibitive marker in Modern Javanese (Ngoko style) *aja* is a recent form of Old Javanese *haywa*. I speculate that the latter derives from *hayu-a, the subjunctive of a static verb root covering the meanings '(be) good, well, right, beautiful, happy'. In analogy with the history of *ora*, it is not unlikely that *hayu-a was originally also preceded by *tan*.

The Modern Javanese Krama form *sampun* followed a different developmental pathway. Originally meaning 'already' and 'finished', it must have become a cessative marker in an intermediate stage before it ended up as a prohibitive (that is, it went from 'finished' to 'stop it!' to 'don't!').

The nominal/contrastive negator *dudu* (Old Javanese *dudū*) must have evolved from a word with a wider meaning that also included 'other, different', if it did not have this meaning in the first place. The association between 'not (this

one, but …another…)' and 'other, different' is also reflected in its Malay counterpart *bukan* and in Proto Western Malayo-Polynesian *bukən/*bəkən 'other, different; negator of nominals' (BTS). The Krama variant *dédé* is derived from *dudu*, conform to a frequent phonological procedure for creating Krama forms. The origin of (Krama/Madya) *sanès* remains obscure.

The origin of the phasal negator *duruŋ* remains uncertain, although formally it must be related to Modern Javanese *wuruŋ* 'fail, failing to happen'. Its Krama counterpart *dèrèŋ* is derived from it according to a phonological procedure for creating Krama forms which was similar to the one deriving *dédé* from *dudu*.

It is clear that a historical investigation of the Javanese negators takes more than a simple application of sound correspondence rules to lexemes that look like one another, whereby the resulting proto forms are given glosses that are basically a semantic compromise distilled from the meanings of the individual reflexes. While the importance of the comparative method is paramount in historical linguistics, it should not make us lose sight of other relevant factors such as morphosyntactic change, word structure, some systematic aspects of irregular sound change, and the considerable value of philology. Philology played a central role in the works of Austronesian comparative-historical linguists more than a century ago but unfortunately has faded into the background since Dempwolff's reconstruction of the PMP phoneme system.

Funding

This publication was supported by the European Regional Development Fund Project "Sinophone Borderlands — Interaction at the Edges" CZ.02.1.01/0.0/0.0/16_019/0000791.

Acknowledgements

When I wrote it I greatly benefited from the expertise of April Almarines, Johan van der Auwera, Arlo Griffiths, Tom Hoogervorst, Anthony Jukes, Willem van der Molen, Elisabeth Riharti, Susi Moeimam, Stuart Robson, Daniel Van Olmen, Jozina Vander Klok, Ljuba Veselinova, and three anonymous reviewers. The usual disclaimers apply.

Abbreviations

Glossing follows the Leipzig Glossing Rules. In addition, the following abbreviations are used:

ANAPH	anaphoric marker	PART	particle
AV	actor voice	PREP	preposition
EMPH	emphatic marker	RDP	reduplication
EXIS	existential marker	TV	transitive
INVOL	involuntative	UV	undergoer voice.
OV	object voice		

References

Adelaar, Alexander. 1992. *Proto-Malayic: The Reconstruction of its Phonology and Parts of its Morphology and Lexicon*. Pacific Linguistics C-119. Canberra: Pacific Linguistics.

Adelaar, Alexander. 1994. "The Classification of the Tamanic Languages (West Kalimantan)." In *Contact-induced Language Change in the Austronesian-speaking Area*, ed. by Tom Dutton, and Derryl Tryon, 1–41. Berlin, New York: De Gruyter Mouton.

Adelaar, Alexander. 1995. "Problems of Definiteness and Ergativity in Embaloh." *Oceanic Linguistics* 34 (2): 375–409.

Adelaar, Alexander. 2011. "Javanese *-aké* and *-aken*: A Short History." *Oceanic Linguistics* 50 (2): 338–350.

Adelaar, Alexander. To appear. "Austronesian Etymologies". In Philip Durkin (ed.), *Handbook of Etymology*. Oxford: OUP, Clarendon.

Aikhenvald, Alexandra. 2010. *Imperatives and Commands*. Oxford: Oxford University Press.

Blust, Robert A. 1979. "Proto-West Malayo-Polynesian Vocatives". *Bijdragen tot de Taal-, Land- en Volkenkunde* 135 (2–3): 205–251.

Blust, Robert A., Stephen Trussel, and Alexander D. Smith. 2023. *CLDF dataset derived from Blust's "Austronesian Comparative Dictionary"*, Vol. 2. Zenodo.

Croft, William. 1991. "The Evolution of Negation." *Journal of Linguistics* 27: 1–39.

Dahl, Östen. 1979. "Typology of Sentence Negation." *Linguistics* 17: 79–106.

Dempwolff, Otto. 1938. *Vergleichende Lautlehre des Austronesischen Wortschatzes. III. Band. Austronesisches Wörterverzeichnis*. Beiheft zur Zeitschrift für Eingeborenen-Sprachen 19. Berlin: Dietrich Reimer Verlag.

Devos, Maud, and Daniel Van Olmen. 2013. "Describing and Explaining the Variation of Bantu Imperatives and Prohibitives." *Studies in Language* 37 (2): 1–57.

Eringa, F. S. 1984. *Soendaas-Nederlands Woordenboek*. (KITLV publication.) Dordrecht and Cinnaminson: Foris Publications.

Ewing, Michael C. 2005. *Grammar and Inference in Conversation: Identifying Clause Structure in Spoken Javanese*. Amsterdam/Philadelphia: John Benjamins.

Gericke, Johann F. C., Taco Roorda, Albert C. Vreede & Johannes G. H. Gunning. 1901. *Javaansch-Nederlandsch Handwoordenboek*. Leiden: Brill.

Gonda, Jan. 1973. *Sanskrit in Indonesia*. New Delhi: International Academy of Indian Culture.

Harris, Alice, and Lyle Campbell. 1995. *Historical Syntax in Cross-linguistic Perspective*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

Hayward, Katharina. 1993. "/p/ vs. /b/ in Javanese: Some Preliminary Data." *School of Oriental and African Studies: Working Papers in Linguistics and Phonetics* 3: 1–33.

Hayward, Katharina. 1995. "/p/ vs. /b/ in Javanese: The Role of the Vocal Folds." *School of Oriental and African Studies: Working Papers in Linguistics and Phonetics* 5: 1–11.

Kern, Hendrik. 1883. "Proeve uit het Oudjavaansche Rāmāyaṇa." In *Bijdragen tot de Taal-, Land- en Volkenkunde van Nederlandsch Indië uitgegeven vanwege het Koninklijk Instituut voor de Taal-, Land- en Volkenkunde van Nederlandsch Indië ter gelegenheid van het Zesde Internationale Congres der Orientalisten te Leiden, Taal en Letterkunde*: 1–25. The Hague: Nijhoff. [Also appeared 1920 in *Verspreide Geschriften* Vol. 9, 251–271. The Hague: Nijhoff].

Kroeger, Paul. 2014. "External Negation in Malay/Indonesian." *Language* 90 (1): 137–184.

Mańczak, Witołd. 1980. "Irregular Sound Change due to Frequency in Latin." *Language Sciences* 2 (1): 62–68.

Miestamo, Matti. 2007. "Negation: An Overview of Typological Research." *Language and Linguistics Compass* 1 (5): 552–670.

Nothofer, Bernd. 1975. *The Reconstruction of Proto-Malayo-Javanic*. Verhandelingen van het KITLV 73. 's-Gravenhage: Martinus Nijhoff.

Ogloblin, Aleksander K. 2005. "Javanese." In *The Austronesian languages of South East Asia and Madagascar*, ed. by Alexander Adelaar, and Nikolaus P. Himmelmann, 590–624. London: Routledge.

Payne, John. R. 1985. "Negation." In *Language Typology and Syntactic Description, Volume I, Clause Structure*, ed. by Timothy Shopen, 197–242. Cambridge University Press.

Pigeaud, Theodoor G. Th. 1938. *Javaans-Nederlands Handwoordenboek*. Groningen: Wolters Noordhoff.

Pigeaud, Theodoor G. Th. 1967. *Literature of Java. I. A synopsis of Javanese Literature 900-1900 A.D*. Leiden: Bibliotheca Universitatis.

Robson, Stuart O., and Singgih Wibisono. 2002. *Javanese-English Dictionary*. Singapore: Periplus.

Sadock, Jerrold M., and Arnold M. Zwicky. 1985. "Speech Acts Distinctions in Syntax." In *Language Typology and Syntactic Description, Volume I, Clause Structure*, ed. by Timothy Shopen, 155–196. Cambridge: Cambridge University Press.

Stevens, Alan M., and A. Ed Schmidgall-Tellings. 2004. *A Comprehensive Indonesian–English Dictionary*, 1st ed. Athens: Ohio University Press.

Uhlenbeck, Eugenius M. 1978. *Studies in Javanese Morphology*. The Hague: Nijhoff.

van der Auwera, Johan, Olga Krasnoukhova, and Frens Vossen. 2022. "Intertwining the Negative Cycles." In *The Negative Existential Cycle*, ed. by Ljuba Veselinova, and Arja Hamari, 611–650. Berlin: Language Science Press.

van der Auwera, Johan, Daniel Van Olmen, and Frens Vossen. 2024. "Negation." In *The Oxford Guide to the Malayo-Polynesian Languages of South East Asia*, ed. by Alexander Adelaar, and Antoinette Schapper, 875–887. Oxford: Clarendon Press.

van der Molen, Willem. 2015. *An Introduction to Old Javanese*. Tokyo: Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Van Olmen, Daniel. 2010. "Typology Meets Usage: The Case of the Prohibitive Infinitive in Dutch." *Folia Linguistica* 44 (2): 471–507.

Vander Klok, Jozina. 2024. "Languages of Java." In *The Oxford Guide to the Malayo-Polynesian Languages of South East Asia*, ed. by Alexander Adelaar, and Antoinette Schapper, 71–487. Oxford: Clarendon Press.

Veselinova, Ljuba. 2014. "The Negative Existential Cycle Revisited." *Linguistics* 52: 1327–89.

Veselinova, Ljuba. 2016. "The Negative Existential Cycle through the Lens of Comparative Data." In *Cyclical Change Continued*, ed. by Elly van Gelderen, 139–187. Amsterdam: John Benjamins.

Veselinova, Ljuba, and Arja Hamari. 2022. "Introducing the Negative Existential Cycle." In *The Negative Existential Cycle*, ed. by Ljuba Veselinova, and Arja Hamari, 1–56. Berlin: Language Science Press.

Veselinova, Ljuba, Leif Asplund, and Jozina Vander Klok. 2024. "Phasal polarity." In *The Oxford Guide to the Malayo-Polynesian Languages of South East Asia*, ed. by Alexander Adelaar, and Antoinette Schapper, 889–906. Oxford: Clarendon Press.

Vossen, Frens, and Johan van der Auwera. 2014. "The Jespersen cycles seen from Austronesian." In *The Diachrony of Negation*, ed. by Maj-Britt Mosegaard Hansen, and Jacqueline Visconti, 47–82. Amsterdam: John Benjamins.

Zoetmulder, Petrus J. 1974. *Kalangwan. A Survey of Old Javanese Literature*. The Hague: Nijhoff.

Zoetmulder, Petrus J. (with the collaboration of S. O. Robson). 1982. *Old Javanese — English dictionary (2 Parts)*. (KITLV publication.) The Hague: Nijhoff.

Zoetmulder, Petrus J. 1983. *De taal van het Adiparwa, een grammaticale studie van het Oudjavaans*. (KITLV Indonesische Herdrukken.) Dordrecht and Cinnaminson: Foris Publications.

Appendix 1. Language ISO-codes

Balinese	ban
Bugis	bug
Chamorro	cha
Embaloh	emb
French	fra
German	deu
Iban	iba
Indonesian	ind
Italian	ita
Javanese	jav
Kambera	xbr

Appendix 1. *(continued)*

Madurese	mad
Malay (macro-language)	msa
Malay (individual language)	zlm
Maloh see Embaloh	
Mentawai	mwv
Old Javanese	kaw
Sanskrit	san
Spanish	esp
Sundanese	sun
Tagalog	tgl
Uma	ppk
Uralic	urj
Zulu	zul

Appendix 2. BTS' (2023) etyma for 'to be' and 'not to be'

Proto Malayo-Polynesian *wada 'be, exist, have; wealthy; not exist, not have'		
Western Malayo-Polynesian		
Maloh [=Embaloh]	ada	don't
Malay	ada	existence; to be, present; to appertain to; (vulg.) to have
	bər-ada	have; rather wealthy, not lacking in means
Mentawai	bara	exist; be present
Old Javanese	ora	not
Javanese	orā [see fn.3]	not (in narrative predicates), no (as a reply, contradiction, etc.)
Madurese	badha	be, have
Balinese	ada	be, exist; existence, being (Low Balinese)
Chamorro	guaha	have, there is, there exists
	gu-in-aha	wealth, affluence, riches, large possessions esp. of worldly estate
Central Malayo Polynesian		
Bimanese	wara	have; wealthy

Appendix 2. *(continued)*

Proto Western Malayo-Polynesian *wada-q 'be, exist, have; not exist; not have'		
WMP		
Itbayaten	wara-wara	things, property, belongings
Ilokano	wadá	to have; to be (somewhere), there is, there are, there was, there were; some
Bontok	wadá	bring about; make; form; create; father a child; there is; there exists
Kankanaey	wadá	there is, there was; to be; be present; exist
Gaddang	wará	there is, there are
Pangasinan	walá	being in existence (often translatable by English 'to have' when referring to possessions, etc.)
Tagalog	walá?	none, nothing; absent
Aklanon	waeá?	disappear, become nothing; run out of; left without
Manobo (Western Bukidnon)	waza?	nothing; without; negative
Sasak	ara?	be, exist, be born
Uma	ara?	there is none
Proto Western Malayo-Polynesian *wada-i 'be, exist, have; not be, not have'		
Western Malayo-Polynesian		
Cebuano	waráy	be gone, not in existence (colloquial)
Iban	adai	born, be born, give birth; supply, provide; have, possess

Note. Also Isneg *wad?óy* 'there is (used in songs)', Bontok *wad?áy* 'bring about; make; form; create; father a child; there is, there exists', Ifugaw *wa* 'there is', Ilokano *addá* 'to have; to be (somewhere), there is, there are, there was, there were; some'. Dempwolff (1934–38) reconstructed ***wada** 'Vorhandensein, Nichtsein'. This apparently contradictory gloss is difficult to improve: the idea of existence or possession is clearly reflected in the northern Philippines, in Iban, Malay, Mentawai, and Chamorro, while a negative sense is clearly reflected in the central and southern Philippines, in Maloh, in Old Javanese and Modern Javanese, and in UMAT of central Sulawesi. These opposing senses do not coexist (without some morphological or lexical modification) in any attested language, and it is unclear how they can be reconciled in a reconstructed gloss. Forms such as Tagalog *walá?* and Cebuano *waráy* appear to be morphologically complex, and show a puzzling parallelism to the morphology of vocatives (Blust 1979).

Epilogue

CHAPTER 32

Epilogue

Episteme and *techne*

Stephen C. Levinson
Max Planck Institute for Psycholinguistics

In the history of science it is often said that the baleful influence of the ancient Greek distinction between *episteme* and *techne*, roughly theoretical science vs. technology and craft, hindered the growth of systematic science until the early modern period. Similarly, the 2025 Nobel Prize in Economics went in part to the historian Joel Mokyr for showing how the failure to integrate practical craft knowledge with scientific theorizing held back economic growth till the beginning of the nineteenth century.

In early modern Europe, the different sciences only slowly and differentially picked up the marriage of technology and science — first astronomy, then physics, then chemistry, then biology and medicine. By the nineteenth century, both the physical world and the biological world were being intimately described, catalogued and analysed. But the language sciences lacked the *techne* for transforming the science of language. On the one hand we had paper and pencil, giving rise to the Boasian trinity of grammar, dictionary and text as the method of encompassing a language new to science. On the other, we had abstract science, the theory of language as Platonic calculus, which dominated the language sciences in the second half of the twentieth century and which, like classical and medieval science had a distain for data collection. Meanwhile, slowly and surely, ever more practical recording devices were becoming available that made it feasible to take recording devices to the field — this was the *techne* whose absence had hindered the great collection of primary data as an analogue to the vast accumulation of information about, for example, botany in the nineteenth century.

It is hard now to imagine the courage it took in the 1990s to stand up to the proponents of linguistics as an abstract science. But Nikolaus Himmelmann had the temerity to point out the absurdity of trying to run before we can walk, that is, to generalize about the general capacity for language before we had the raw data for many languages. In his now famous paper of 1998 he outlined how such data collection might be best done, foreseeing many of the parameters we now take for granted (ethical issues, naturalness issues, issues of representative-

https://doi.org/10.1075/slcs.240.32lev
Available under the CC BY-NC-ND 4.0 license.
© 2026 John Benjamins Publishing Company

ness), so laying the foundation for the field he called *language documentation*. The timing was perfect. First, the *techne* had caught up with the aspirations — we now had digitized video recordings, making possible something like living documents of speakers in action. Second, we had in the shape of the Volkswagen Foundation a funder for what became the DobeS Project with an enthusiastic group of scholars, both senior and junior. Third, we had some visionary technicians, notably Peter Wittenburg at the Max Planck Institute for Psycholinguistics, who were minded to bend the still nascent technology to the documentary tasks. So, language documentation in its modern incarnation was born, not without many birthing pains, it should be said: disagreements about aims and methods, impossibly clumsy early technologies and growing realizations about ethical problems.

Where are we now? Himmelmann's vision of a primary documentation, with video records of all major genres, is up and running. It's a respectable calling, made reputable by the interdigitation of theory and practice along the lines Himmelmann called for, so that a documentary linguist is now a desideratum in any modern linguistics department. It is possible at last for a linguist to inspect the primary recordings from an endangered or extinct language, just like a botanist has been able for some two hundred years to go to an herbarium and inspect a dried specimen from a rare or extinct plant. Native speakers now apply for funding, and through, for example, the Endangered Languages Documentation Programme have a chance to contribute to the documentation of their own languages. But there are disappointments. Unfortunately, unlike the modern botanical seed banks, language revitalization remains a major challenge. It also remains incredibly labour intensive to transcribe, annotate and translate these primary records, and so get them into a state where secondary description and analysis can profitably be done. These annotated corpora are therefore small, and cover limited phenomena. It remains hard to get the sustained funding to provide rich annotation and rich metadata for large collections. If we can hold out, artificial intelligence (that is, machine learning) may well come to help us in both the annotation and the extraction of grammatical and semantic patterns by learning to extrapolate from small corpora. Another serious hurdle is that no funding agency likes archives — who would want to fund a never-ending commitment, a continual fight to hold the *status quo* through technology upheavals, when instead they could go on exciting adventure holidays in scientific no-man's land? Only the Arcadia foundation, the Max Planck system, CNRS, and a few national libraries or university consortia (e.g. PARADISEC, AILLA, Language Archive Cologne) currently carry the burden.

Meanwhile, we have leap-frogged over the primary records, and attempted to make large summary data bases of syntactic typology and phonology. I myself have played a role in the Glottobank enterprise now spear-headed by the Max

Planck Institute for Evolutionary Anthropology, viewing it as a necessary stopgap — these data bases extracted from records more in the Boasian style are what allows the little discipline of linguistics to play with the big boys, with archaeology, genetics, history, psychology and biology. We need to be in that larger game if the language sciences are to survive — for when the Chomskyan vision faded, philosophy, psychology and neuroscience largely turned away from language. These data bases, we can hope, will buy us the time to first, increase the range and size of our primary data archives, and second, learn how to use machine-learning to help extract the grammatical, semantic and lexical patterns out of them. The vision is to be able, at the click of a button, to get from theoretical statement to sets of primary records, and from primary records to clear generalizations. Then at last, the language sciences would be on the firm footing that all of us now hope for, and Nikolaus first dreamed of.

Language index

A
Adamorobe Sign Language 194
Aklanon 897
Ambon Malay *See* Malay
American Sign Language 194–196
Amis 544–546
Ancash Quechua *See* Quetchua
Ancient Egyptian 239, 241–242, 251, 256
Apatani 533
Arabic 583
  Egyptian Arabic 401–402
  Lebanese Arabic 402
  Moroccan Arabic 405
  (West) African Arabic 271
Arta 500
Aru Malay *See* Malay
Australian languages 117, 122–123, 126, 297–299, 481, 789, 801, 805, 808–809
Australian Sign Language 218
Austronesian languages 29, 45–51, 64, 75, 292–293, 350, 380, 500, 518–524, 540, 544–545, 550, 561–563, 572, 610, 620, 695 *See also* western Austronesian
Awiakay 316–323, 329–332, 335–337, 339

B
Bacan Malay *See* Malay
Balinese 75–77, 80, 82–83, 92–95, 101–102, 579, 616, 619, 886, 895–896
Banda Malay *See* Malay
Besemah 572–572, 576–606, 619
Biak 50–51, 54–55, 856, 859–860, 862
Bikol 553
Bimanese 896
Bininj Gunwok *See* Bininj Kunwok
Bininj Gun–wok *See* Bininj Kunwok
Bininj Kunwok 110, 116–118, 123, 125, 128, 295, 402, 789, 791, 797, 803–804, 807–810
Bontok 897
Bora 379–380, 383, 390, 396, 500
Brazilian Portuguese *See Portuguese*
Brazilian Sign Language 194
British Sign Language 195, 218, 222
Brunei Malay *See* Malay
Budai Rukai 590
Bugis 886
Bugkalot 556–557
Bunun (Takituduh) 544–545
Burarra 810

C
Cahita languages 524, 534, 538, 541–544, 550 *See also* Old Cahita
Cebuano 552, 558, 562, 897
Chamorro 532, 560, 874, 896–897
Cha'palaa 764
Chipaya 724–726, 733, 741, 743–747, 749–750
Classical Malay *See* Malay
Chirag 500
Cuzco Quechua *See* Quechua

D
Dalabon 107, 109–110, 112–129, 402
Dene languages 351
Dëne Sųłiné 343–351, 362–368
DGS *See* German Sign Language
Dolakha Newari 526
Dutch 220, 390, 402, 494–495, 583

E
Eipo 57, 61, 86, 315–316, 318–321, 323–325, 329, 332–334, 338
Embaloh 874–875, 886, 896
Enggano 31, 75–78, 83, 91, 95–100
English: Modern 219–220, 222, 380–381, 383, 385, 390, 396, 398–400, 402–405, 500, 503, 507, 624, 766, 796, 802
  Middle 339, 765–766
European Portuguese *See* Portuguese

F
French 167, 220, 380, 383, 385, 396–398, 400–402, 407–411, 414–415, 417–420, 876 *See also* Swiss French
French Sign Language 195

G
Gaddang 897
German: New High 219–222, 225–226, 389, 397, 399–402, 407–412, 414–420, 715, 717, 757, 766, 768–771 *See also* Swiss German
  Middle High 765
  Middle Low 765–766
German Sign Language 221, 224
Greek 397, 398, 400–402, 407–411, 414–420
  Cypriot Greek 500
Gunwinyguan languages 109, 116, 118, 129, 786, 789, 802–805, 808–810

H
Huallaga Quechua *See* Quechua
Hawai'i Sign Language 196

I
Iban 786–787, 874–875, 897
Igu 287, 300, 302, 304–305, 307
Iha 50, 55, 57, 61, 69, 86, 343–357, 364–368
Imbabura Quechua *See* Quechua

Indonesian 52, 90, 99, 560–562, 575–577, 584, 592, 601, 603, 610, 615–616, 619, 622, 627, 854, 870–871, 890 *See also* Jakarta Indonesian
Indonesian (type) languages 94, 574–576, 578, 582–583, 587, 591–592, 601, 610, 612–613, 615–619, 629, 634–635
Indo-European languages 222, 290–294, 305, 546, 550, 750, 763, 766
Italian 220, 398, 402–404, 444
Itbayaten 897

J
Jakarta Indonesian 610, 612–613, 615, 619–624, 627–628, 631–635
Jaminjung 671, 673–697
Japhug 531–532
Javanese 575, 853, 866–892, 896–897
Jinghpaw 500

K
Kalamang 55, 500
Kambera 886
Kanakanavu 559
Kankanaey 897
Karachay-Balkar 526–527
Kata Kolok 194
Kayardild 117, 128, 791, 807–809, 829
Kera'a 287, 296–300, 302, 304, 307
Korean 547 *See also* Soul Korean
Kupang Malay *See* Malay
Kunbarlang 129, 786–811, 831

L
Lardil 807, 809
Libras *See* Brazilian Sign Language
Lithuanian 763, 767, 771–773, 775–778
LSF *See* French Sign Language

M
Madurese 576, 611–612, 874, 896
Malay 547, 560–561, 575, 579–580, 582–583, 587, 591, 601, 603–604, 845, 853, 856–857, 862–863, 870–871, 874–875, 890, 892, 896–897
  Ambon Malay 845–852, 855–856, 862
  Aru Malay 846
  Bacan Malay 847–852, 561
  Banda Malay 855
  Brunei Malay 886–887
  Classical Malay 619, 855
  Kupang Malay 856
  Manado Malay 847–852, 854–855, 858, 861
  North Moluccan Malay 845, 847–852, 854–856, 858–862
  Papuan Malay 49, 52, 55, 844–863
  Sarawak Malay 886–887
  South Barisan Malay 573, 579
Maloh *See* Embaloh
Manado Malay *See* Malay
Mandarin 402, 500, 503, 507
Manobo 554–556, 558, 897
Mardin Sign Language 196
Marind 49–50, 830
Marori 50, 55, 57, 75, 78, 82, 90, 95–100, 102
Matukar Panau 500–501
Meakambut 315–317, 319, 321, 339
Mekwei 860
Mentawai 874, 896–897
Meto 429–430, 436–437, 439–458
Miraya Bikol *See* Bikol
Mongsen Ao 525, 543
Movima 494–495, 524, 644–668

N
Nafsan 380, 500
Ngaliwurru 671, 673–697
Nicaraguan Sign Language 194
non-Austronesian languages 29, 45, 350, 541 *See also* Papuan languages
North Moluccan Malay *See* Malay
Northern Kurdish 380, 383, 396, 500

O
Old Cahita 535–536, 538, 541–542 *See also* Cahita languages

P
Paiwan 522, 544, 551
Pangasinan 897
Papuan languages 46–49, 64, 75, 350, 815–816, 820, 823, 835, 839, 863 *See also* non-Austronesian languages
Papuan Malay *See* Malay
Persian 193, 500, 504, 506
Philippine (type) languages 518–520, 524, 533, 546, 550–554, 557–563, 610, 612–619, 629, 634–635
Pima Bajo 525
Pitta–Pitta 805–807
Plains Indian Sign Language 218
Portuguese: Brazilian Portuguese 494, 495, 504, 768
  European Portuguese 397–398, 400–402, 407–411, 414–420, 768, 853, 856, 858
Puyuma 519, 522–524, 532, 538–541, 559, 561, 563

Q
Qaqet 275, 343–353, 359–368, 435
Quechua: Ancash Quechua 528
  Cuzco Quechua 529, 548–550
  Huallaga Quechua 528
  Imbabura Quechua 548

R
Ratahan 617
Romani 265, 272, 275–279
Rongga 75–76, 78
Russian Sign Language (RSL) 218, 221–222

S
Saaroa 540
Samoan 401–402, 724–726, 732, 734–737, 740–741, 749–750
Sanskrit 881
  Vedic Sanskrit 300, 763
Sanzhi Dargwa 380, 383, 500

Sarawak Malay *See* Malay
Sasak 586, 597
Seediq (Truku) 562
Sinasina Sign Language 195
Siraya 544, 562
Slovenian 765
Soul Korean 397–398, 402, 407–411, 414–415, 417–420, 427–428
South Barisan Malay *See* Malay
Spanish 219–220, 398, 401–405, 496, 498, 853
Sundanese 874, 890
Sumbawa 500, 873
Swiss French 399
Swiss German 399

T
Tabasaran 380, 383, 396, 500
Tagalog 489, 507, 519, 522–523, 546, 550–553, 610, 611–615, 619–627, 629–630, 632–635
Tahitian 464–468, 470–482
Tausug 562–563
Teop 276, 380, 383, 396, 500
Ternate 844, 846–855, 857–863
Thao 544–545
Tibetan 524, 530–531, 541, 547
TİD *See* Turkish Sign language
Tondano 500, 619
Totoli 381, 560, 574, 579, 610, 612–613, 616–622, 624–625, 628–629, 632, 634–636
Tulil 500
Turkish 444, 526–529
Turkish Sign Language 196

U
Uma 875, 897
Uralic 292–293, 380, 432, 878
Uruangnirin 500
Ute 526, 529–530, 534

V
Vera'a 380, 396, 494, 496, 498, 500

W
Waray 554
Warlpiri 802, 806
Warlpiri Sign Language 194
Western Austronesian 611–614, 619–620, 632, 636
Western Bukidnon Manobo *See* Manobo
Wooi 49, 55–61, 69, 86, 90

Y
Yali 50, 52, 55, 57, 61–62, 69, 86, 815–839
Yandruwandha 805–807
Yale 57, 61, 86, 314–316, 318–320, 323, 328–329, 332, 338
Yaqui 534–535, 538, 541–542
Yolngu Sign Language 194, 218
Yucatec 758–759
Yurakaré 222, 762, 766, 768–770

Z
Zulu 882–883

Subject index

A

acceptability (judgement) 714–720
archives and archiving 2–3, 9–10, 16–18, 37–43, 55–57, 59–61, 67–68, 100, 108, 110, 126, 130, 134–142, 165, 184, 186, 198–199, 216, 221–224, 228, 280, 409, 437, 466, 580, 621, 646, 685, 687, 697, 769, 816, 901–902

C

capacity building 43, 59–60, 62–63, 75, 77, 79, 86–87, 91–92, 96, 98–102
community-led documentation 75–76
conversation (analysis) 107–108, 111, 123–127, 137, 185–189, 200, 213–214, 216, 218, 220–223, 225–228, 308, 319, 376, 495–496, 503, 505, 520, 572, 576–577, 579–580, 584–586, 588–591, 595, 598, 603, 605, 685, 756–778, 790, 797, 805, 816–817, 828
corpus building 210–228
corpus linguistics 219, 376, 392
child language 276, 346, 610–636

D

decolonial linguistics 76, 89, 99, 309
diachrony 19, 77, 167, 223, 287–309, 350, 377, 489, 493, 506, 520, 531, 538, 541, 543, 545, 615, 756–778, 802, 805, 808–809, 819
dialect 36, 51, 54, 62, 84, 93, 96, 137, 166, 265, 278, 289, 295–296, 304, 307, 316–317, 321, 328–329, 339, 348, 365, 402, 429, 436, 519, 528, 552, 561, 573, 579, 621, 767, 772, 774, 776–777, 803, 868, 878, 884–885
dialogue 89, 126, 129, 288, 303, 586, 679, 719
dictionary (writing) 62, 77, 83–85, 109, 118–119, 121–125, 166, 318, 349, 446, 807, 810, 855, 868, 875–876, 900
discourse (analysis) 5–6, 14, 37, 55, 80, 116, 127, 156–157, 160, 168, 173–175, 185, 187, 221, 243–246, 249, 252, 265, 268, 300, 306, 344, 489–491, 494–500, 503, 505–508, 572–573, 579, 584, 587, 591, 593–594, 603, 646, 653, 655, 657–668, 672–673, 675–676, 678, 681–682, 685, 687–688, 690–691, 694–695, 705, 707–708, 710–713, 720, 724–750, 761, 766, 797, 815, 822

E

endangered languages 2, 9, 11, 15–16, 18, 30–31, 33–36, 47, 51, 55, 57–58, 68, 76–77, 85–88, 108, 137–139, 151, 163–166, 170, 196, 198, 215–216, 219, 223, 266, 268–270, 274, 290, 296, 299, 305, 344, 349, 378, 464–465, 580, 612, 620, 646, 769, 789, 901
ethical aspects of language documentation 3, 15, 28, 38, 40–42, 76, 79, 84, 88–89, 99–100, 102, 136, 140–142, 195, 280, 289, 307, 685, 900–901

F

folk story 300, 305, 308, 346, 743, 761

G

gender 84, 153, 349, 409, 498, 504–505, 728
genre 187, 241, 245, 251, 256, 270, 288–289, 291, 299–302, 304, 306–309, 351, 321, 378, 464, 471, 481–482, 696, 725, 749, 761, 778, 820, 889, 901
gesture 183, 185–186, 188, 190–193, 195–196, 198–201, 212–213, 219–221, 223–224, 227, 239–240, 242–244, 247, 249, 251–256, 268, 492, 707, 720
grammar (writing) 3–4, 36, 54–56, 76–80, 82–85, 93–94, 166, 189, 192, 201, 216, 295, 318, 339, 346, 494, 551
grammatic(al)ization 5–6, 290, 294, 518, 537, 704, 710, 791, 796, 808, 815–839

H

historical (linguistics) 14, 39–40, 55, 127, 195, 223, 242, 256, 266, 272, 287–309, 351, 376–378, 498, 520, 522, 545–547, 550, 556, 559, 580, 756–778, 805, 809–810, 817, 845–846, 854–857, 860–863, 866–867, 869, 871, 874, 878–879, 881, 883, 889, 891–892

I

information packaging 242, 250, 696
information structure 193, 481, 671–696
infrastructure (incl. funding infrastructure) 2, 12, 16–18, 54, 58, 60, 134–142, 155, 167
interaction(al linguistics) 1, 79, 82, 126, 183, 185–186, 188–190, 197–202, 210, 212–213, 215, 218–228, 243–244, 251, 269, 272, 304, 308, 319, 346, 385, 576, 587, 605, 621–622, 674, 685, 708, 710, 714, 719, 727, 758, 761, 768–770

L

language acquisition 610–636

language change 288, 291, 306, 376, 671, 756–778, 818
language contact 46–47, 80, 166, 187, 194, 199, 271, 289, 292, 295–296, 299, 316, 845–846, 857–859, 861–863
(language) policy 10, 42, 140, 151, 153, 163, 173, 279
lexicon 3, 60, 78, 129, 216, 436, 808, 847, 854, 862–863, 869
lifespan 289
literacy 54, 78, 84, 263–281, 290, 300, 303, 305–306, 348, 368

M
metalinguistic knowledge 83, 122, 217, 275–277, 280, 343–368, 706, 714, 719, 816–817, 826, 870
morphology 4, 48–49, 55, 63, 112, 120, 162, 290, 294, 296, 302, 362, 410, 432, 434, 449–450, 452, 454, 457–458, 489, 491, 499, 505–507, 519–520, 522, 524–525, 529, 532, 535, 538, 544–546, 550, 551, 553, 555–556, 560, 562, 563, 612, 614–615, 617–620, 622–623, 625, 627–629, 631–632, 635, 645–646, 648–650, 653, 654, 657, 666, 672, 674–677, 685, 745, 790, 804, 809, 820, 823, 828, 868, 873, 881, 897
multilingualism (incl. bilingualism) 14, 51–53, 78, 80, 83, 89, 108, 167, 196, 269, 273, 289, 471, 489, 499, 621, 769
multimodality 79, 173, 182–183, 185, 187–190, 192–195, 197, 200–202, 210–228, 239–256, 308

N
narrative 3, 32, 89, 135, 185–187, 199, 215–216, 219, 268, 279, 300, 302–303, 305, 321, 379, 437, 471, 489, 494, 499, 503, 505, 508, 659, 683, 685, 724–750, 761, 769, 789–790, 798, 799, 816–817, 826, 896

O
orality 30, 42, 126, 164, 265, 267–269, 272, 288–289, 299–309, 343, 499, 646, 724–725, 732–733, 745, 749–750
orthography 78, 187, 199, 266–267, 270, 273, 275, 277–279, 343, 345, 347–350, 353–355, 359, 362–368, 385, 458, 471, 649, 790, 867

P
phonetics 14, 31, 137, 239, 297, 398, 344–345, 350, 352–354, 359, 361–367, 376, 429–458, 464–482, 817, 889
phonology 55, 63, 122, 162, 191, 199, 275, 290, 294, 297, 299, 314, 345, 351–352, 362, 365–367, 385, 429–458, 464–482, 491, 494, 504–505, 508, 544, 674, 809, 819, 854–855, 860, 875, 877, 879, 881, 884–886, 891–892, 901
positionality 75–76, 88, 97–98
prominence 7, 14, 138, 158–159, 168, 172, 239, 241, 247, 249, 256, 294, 400–402, 406, 408, 415, 419, 421, 467–469, 475, 477–478, 481, 489, 491–492, 501, 506, 520, 593, 662–664, 733, 868, 874
prosody 6, 7, 49, 108, 185, 190, 199, 221, 225, 348, 376–377, 391, 397–398, 400, 402–403, 406–408, 419–422, 431, 445, 448, 464–482, 489, 506, 538, 649–650, 655, 661–664, 672–673, 675, 678, 680–686, 690–692, 695, 720, 727, 817–818, 832–833, 836, 838

R
rhetorical structure 301, 724–750
ritual language 89, 194, 242, 244, 265, 288, 300–302, 307, 315, 321, 332, 761

S
science communication 18–19, 140–175
script 128, 247, 265, 271, 273, 275, 278
semantics 55, 63, 80, 94, 96, 107–109, 111, 117–120, 122, 128, 130, 158, 162, 170, 190, 240, 246–247, 249–250, 294, 304–305, 314–315, 449, 504, 530, 533, 546, 558, 597, 605, 611, 614, 617, 622–623, 647, 654, 665–666, 668, 676–677, 720, 725–728, 731–732, 737, 741, 745–746, 748–749, 767, 788, 791, 795–796, 802–804, 806–809, 811, 815, 817–819, 822, 824, 826–835, 837–838, 854–855, 863, 875, 881–883, 885–886, 889–892, 901–902
sign language 14, 185, 188–190, 194–196, 201, 210–228
sociolinguistic(s) 33, 47, 51, 53–55, 63, 84, 127, 188, 289–290, 298, 580, 620–621
spoken language 3, 5–7, 19, 31, 37, 187–188, 195–196, 201, 211–212, 214, 221, 240, 244, 251, 256, 265, 268–269, 271–272, 276, 279–280, 308, 314, 347–348, 392, 489, 497, 499, 673, 684–685, 727, 764, 817, 838–839, 874, 881
spontaneous speech 55, 306, 308, 381, 481, 625–626, 633–634, 816, 839
stimulus 79, 127, 301, 399, 759, 816
syntax 3–6, 55, 63, 78, 80, 83, 93–94, 276, 294, 305–306, 322, 376, 452, 489, 491–494, 496–499, 508, 519–520, 524, 550, 557, 559–563, 573–575, 579–580, 588, 593, 644–649, 651, 655–657, 662–663, 665–668, 672, 675–677, 685, 734, 743, 788–790, 792–794, 796, 810, 818–819, 826, 829, 831, 833–834, 836, 838, 883–884, 892, 901

T
transcription 7, 83, 96, 108–110, 115–117, 119–120, 122, 129–130, 164, 187, 197, 199–200, 217–218, 248, 268–269, 275, 277–279, 281, 338, 343–368, 378, 385, 410,

458, 499, 573, 621, 685, 735, 743, 788, 790, 801, 817, 823
typology 1, 4, 19, 78–80, 82, 85, 87, 93–94, 102, 108, 139, 154, 183, 193, 196, 214, 292–297, 299, 305–306, 318, 376, 398, 400, 405–406, 420, 422, 447, 465–466, 469, 480, 489, 495, 506, 508, 518–519, 537, 563, 610–613, 616–617, 618, 634, 672, 768, 831, 835, 875, 882

W
writing 39, 95–96, 102, 163, 170, 175, 217, 239–241, 244–245, 256, 263–281, 290, 301, 303, 305–306, 345, 348–349, 353, 360–362, 265–368, 348, 732, 773–777, 884

The documentarist turn, i.e., the growing impact of documentary concepts and practices within general linguistics, has increased awareness of the empirical foundations of our discipline, accompanied by an appreciation of the value of observable linguistic behaviour. Today, there exist well-annotated, audiovisual data collections for many – if still far from enough – endangered or un(der)described languages across the world, each striving to present a "lasting, multipurpose record of a language" (Himmelmann 2006). Documentary data is increasingly supporting theory-building on a larger scale, which is where the present volume steps in: What does the shift towards more diverse, ecologically valid data types mean for synchronic and diachronic linguistic analyses and typological generalizations? This question is addressed by this volume's far-ranging contributions on academic and societal affordances (Part I), interrelations between corpus design and avenues of research (Part II), as well as synchronic (Part III) and diachronic (Part IV) research questions.

ISBN 978 90 272 3400 1
9 789027 234001

JOHN BENJAMINS PUBLISHING COMPANY