A Digital Reference Grammar of Abesabesi

Towards a data format for digital reference grammars

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophischen Fakultät der
Universität zu Köln
im Fach Linguistik

vorgelegt von

Jonas Lau
geb. am 04.09.1990
in Hilden

Köln, 23.03.2021

Eidesstattliche Erklärung

Ich versichere eidesstattlich, dass ich die von mir vorgelegte Dissertation selbständig und ohne unzulässige Hilfe angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Dissertation einschließlich Tabellen, Karten und Abbildungen, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Hochschule zur Prüfung vorgelegen hat; dass sie, gegebenenfalls abgesehen von einer durch die Vorsitzende oder den Vorsitzenden des Promotionsausschusses nach Rücksprache mit der betreuenden Hochschullehrerin beziehungsweise dem betreuenden Hochschullehrer vorab genehmigten Teilpublikation, noch nicht veröffentlicht worden ist sowie dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen in §§ 20 und 21 der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Birgit Hellwig, Prof. Dr. Andreas Witt und Prof. Dr. Nikolaus Himmelmann betreut worden.

Köln,

Jonas Lau

Acknowledgments

This PhD project, including this thesis, the fieldwork, and the development of the digital reference grammar, has not only been the result of many hours of work but also of generous amounts of help and advice by multiple people I would like to thank.

I am most grateful for the people of Ikaram, who welcomed and accomodated me in their town, taught me about their language and culture, answered my questions, and took care of me. I thank H.R.M. Andrew Momodu, the Akala of Ikaram, for accepting me at the palace and providing me with a temporary home, for a lot of advice, and for presenting me to the people of Ikaram. I thank Dr. Taiwo Agoyi for being my first anchor in the Akoko hills, for inspiring my research, for presenting me to all the great researchers and students at Adekunle Ajasin University in Akungba, for bringing me to Ikaram, and for all the moral support during my field trips. I thank Olori Adenike and Olori Ilile for taking such good care of me, providing me with ever delicious meals, and making the palace feel like home. I thank Arohunmolase Haruna for taking care of my security, showing me all the nice places of Ikaram, and climbing the hills with me. I thank Steven Ayodele Awolami, Chief Abiodun Afolabi, Chief Alice Elegbeleye, and the Eleperi Chief Olaniran Samuel Jimo for your never-ending wealth of stories, for all the songs and proverbs you taught me, and for all the help you provided for my research. Thank you Laisi Rafiu for translating stories and narratives for me, for being patient with me, and for putting up with all of my questions. I also thank Mama Joy for providing me with biscuits, snacks, and anything else I needed, and for teaching me bits of your language. And I am grateful for all the other speakers that told me stories, translated sentences, and helped me with my questions. Their names are Janet Akeju, Bashiru Naibi Aminu, Caroline Elegbeleye, Prince Olubunmi Victor Daudu, Medinat Folashade Oludade, Adeolu Mercy Funke, Alamo Idris Alooma, Shittu Samuel Aderemi, Ajatta Adebowale Ilesanmi, Friday Jimo Ibrahim, Weehab Aminu Oshono, Victoria Aderinsola Ayetoro, and Kayode Seth Banfe. I am truely grateful to all of you!

I also want to thank everybody that supported me at the University of Cologne. First of all, my first supervisor Birgit Hellwig, who was always there when I needed feedback, support, or any advice from funding to academic writing to fieldwork methods. Thank you for all your effort and enthusiasm. I could not have imagined a better first supervisor! Many thanks also to my second supervisor Andreas Witt, who directed me on my path through the digital humanities and gave valuable advice and feedback on the development of this grammar, its data format, and this thesis as a whole. I would also

like to thank my third supervisor Nikolaus Himmelmann, and my former office mates Felix Rau and Luiz Antonio de Sousa Netto. Thank you for inspiring thoughts, a comfortable working atmosphere, and all your feedback. Lena Wolberg has helped me out in numerous situations: organizing my finances, ordering my equipment, and so much more. Thank you for your support, your efficiency, your organizing skills, and all the nice conversations!

For all support regarding TEI, thanks to Toma Tasovac and Laurent Romary, who taught at the Lexical Data Masterclass and who have continued giving me feedback and advice on my data format. For technical support, I would like to thank the Data Center for the Humanities (DCH) and the Cologne Center for eHumanities (CCeH) – especially Patrick Helling, Sviatoslav Drach, and Jonathan Blumtritt. Thank you for helping me set up and host the application and for providing me with support for eXist DB.

I am also very grateful for the financial support I received to make this research possible. I was granted a PhD scholarship from the a.r.t.e.s. Graduate School for the Humanities Cologne and benefited from their support and infrastructure. I received valuable feedback and inspiration from my graduate class, supervised by Martin Becker and Nicolas Pethes. And I was able to carry out my fieldwork with a grant from the Endangered Languages Documentation Programme (ELDP). I am grateful for the programme's support and for the opportunity to deposit my research data in the Endangered Languages Archive (ELAR). A very special thank you goes to Sophie Salffner, who was my first signpost to Ikaram. I am truely grateful that you suggested that I work on Abesabesi, for getting me in contact with Taiwo, for providing lots of inspiration through your archive deposits and research, and for all your feedback and support. I would not have gotten to know Ikaram without you!

A big thank you goes to my dear friend Sarah Dopierala, who made all my sentences beautiful and gave me tons of feedback, inspiration, and support throughout my PhD. Ich danke meiner Familie und meinen Freunden, die mich durch die letzten Jahre von den Bewerbungen bis zur Abgabe begleitet und mir Kraft gegeben haben. Und zu guter letzt gilt ein großes Dankeschön Jonas, der mir Halt gegeben hat, mich durch alle Ängste und Freuden begleitet hat, und mich glücklich macht.

Contents

Chapter 1

Introduction

Digital technologies have not only transformed many aspects of private lives and the global economy, they have also made their entry into different fields of research. Within the discipline of linguistics, the field of language documentation and description in particular has benefited greatly from digital data formats and software to manage its vast amounts of data. Some digital formats are used extensively and have almost replaced their analog counterparts. Language documentation, for example, relies on digital data formats for dictionaries, time-aligned annotations, and metadata. For some analog formats, there are no digital counterparts. For others, digital formats have been proposed, but they have not yet overcome the restrictions posed by their analog counterparts. One such format is the reference grammar. As it is tied to the rather conservative domain of publishing, it needs to follow the standards provided and demanded by publishers. Digital publishing exists, but digitally published reference grammars rather resemble a digital copy of the paper grammar. Instead, a digital format specifically designed for reference grammars could overcome the limits of a paper grammar and add new possibilities.

Although authors and readers could both strongly profit from a digital reference grammar (DRG), the development of a standardized format with accompanying software seems to entail various hurdles. Such an undertaking would require more than a single researcher and a generous amount of funding (Drude, 2012, p. 161). This thesis, therefore, does not aim at the development of a general grammar platform where grammars can be authored and explored. It rather develops a DRG on a small scale and attempts to generalize the experiences of the author for a wider context of use whenever possible. Within this scope, the foundations toward a digital grammar format will be laid, which could be used by a large scale grammar platform project in the future.

A great number of reference grammars are produced every year within the scope of documentation projects. The field of language documentation and description is thus a major producer of reference grammars. At the same time, a variety of language structures are represented since language documentation is concerned with un- or under-described languages all over the world. The produced reference grammars are mainly directed at an academic audience of linguists interested in typology, the language family, the linguistic

area, or the language in particular.

The developed grammar will be a DRG of Abesabesi[1], an endangered Benue-Congo language of South-West Nigeria. Abesabesi is spoken by approximately 7000 speakers in nine settlements of Ondo State. This PhD project encompasses a small documentation project funded by the Endangered Languages Documentation Programme (ELDP), where data has been collected that will form the base of the grammar. In order to remain within the limits of a PhD project, this grammatical description will be compiled in the form of a sketch grammar. The proposed grammar format will be used to encode the Abesabesi Sketch Grammar. Moreover, an application is developed to facilitate the exploration of the grammar. The application is able to read the grammar format and display its content in a reader-friendly way by offering a range of enhanced functionalities such as navigation tools and multimedia embedding. As applications would have to be maintained and, therefore, be backed by a larger institution, this thesis focuses on the development of a grammar format which could be used or extended by other projects. It uses the existing base for text encoding – the Text Encoding Initiative (TEI) – and proposes a way to use this standard for encoding reference grammars. Moreover, a detailed proposal on how to encode Interlinear Glossed Text (IGT) in TEI is presented.

1.1 Research questions and outline

This thesis attempts to answer multiple questions concerning the structure of the language Abesabesi as well as the benefits and feasibility of a DRG. Part I of this thesis will pertain to the language Abesabesi itself. Part II is then concerned with grammaticography and digital grammaticography in particular.

Part I Chapter 2 introduces the language Abesabesi, its socio-linguistic context, and the documentation project. After that, Chapter 3 contains the Abesabesi Sketch Grammar. It presents the findings of the documentation project and provides the reader with the content of the Digital Reference Grammar of Abesabesi. This information will later be used in discussions on encoding the grammar. It is important to note, however, that I did not strive to retro-digitize the linear sketch grammar presented in Chapter 3. Therefore, Chapter 3 and the encoding of the DRG were produced simultaneously. This enabled me to contrast the two structures but might have also influenced my style of encoding.

Part II As the general aim of this part is to conceptualize a digital alternative to paper grammars, it deals with the question “What could a digital alternative to paper grammars look like?”. Chapter 4 provides a literature review about grammaticography and digital grammaticography. It introduces the concept “grammar”, reviews existing research on

[1] ISO-639-3: ibe, Glottocode: akpe1248. The developed grammar application can be accessed at `https://abesabesi.cceh.uni-koeln.de/`

different types of grammars, their structure, and content. Moreover, it reviews different approaches to digital grammaticography and discusses the notion of a DRG. Chapter 5 then contrasts the two manifestations of reference grammars – digital and analog. It deals with the questions "Could reference grammars benefit from a digital format?", "Which properties of paper reference grammars can be transferred to digital reference grammars?", "Which new properties and functionalities of digital reference grammars distinguish them from paper reference grammars?", and "What are the limits of digital reference grammars?". Although paper reference grammars are valued for multiple things, digital reference grammars could open up new possibilities for authors (such as guidance for structuring the grammar) and readers (such as full text search). The last section of Chapter 5 summarizes the chapter by conceptualizing a DRG and its structure.

The structure of a reference grammar in general is analyzed in Chapter 6 in order to transfer it into a digital format. The chapter thus deals with the questions "What are common elements of reference grammars?", "How can reference grammars be modeled?", "How could a digital reference grammar be stored?" and "What could a digital format for reference grammars look like?" At the end of the chapter, a format for DRGs is proposed.

In order to test the format proposed in Chapter 6 and to visualize the encoded Abesabesi Sketch Grammar, a grammar application is developed. This application functions as an exploration tool and reads the digital grammar format. The chapter discusses its development and the implications of DRGs and grammar platforms on a larger scale.

The final Chapter 8 presents recommendations for a project on a larger scale. It uses the experience of the author and discusses how this experience can be used as a reference for the development of future projects.

1.2 Linguistic conventions

All Abesabesi examples are transcribed according to a practical orthography discussed in Section 3.1.6. In-text examples are formatted in italics to distinguish them from the surrounding text. They might be accompanied by a gloss or translation, which will be enclosed in inverted commas: *odʒíbèrɛ̀* 'gorilla'. More complex examples are presented as interlinear glossed text in three tiers. The first tier presents the text where morphemes are separated by hypens. The second tier provides glosses for each morpheme, and the third tier contains a free translation. In some cases, an additional tier on the top indicates syntactic functions. These interlinear examples reference the original utterance in the corpus (Section 2.3.3) by adding a segment identifier in parentheses to the right of the example (Example 1.1)

(1.1)	*S*	*V*	*PO*	*V*	*PO*		
	ǹ	ton	u	kèd	ɔban	u	
	1SG	meet	3SG.OBJ	carry	child	3SG.POSS	
	'I helped her to carry her child'						(ibe073-00.173)

Segment identifiers refer to a specific segment in a recording within the archive deposit. The deposit can be accessed through the following link: `http://hdl.handle.net/2196/00-0000-0000-0012-6E05-F`. The corresponding session can be found using the first six characters of the identifer (ibe073). Within the session, the corresponding recording/annotation file is indicated by the first nine characters (ibe073-00). While audio recordings have the extension `.wav` and video recordings have the extension `.mp4`, annotation files have the extension `.eaf`. The last three numbers of the identifier indicate the segment number within the corresponding annotation file (173). Session ibe001 of the text corpus refers to a recording made by Sophie Salffner and cannot be found in the Abesabesi corpus. The session can be found in Salffner's deposit at `http://hdl.handle.net/2196/eaacb80f-8877-4cea-8573-63c9f526393b`.

The glossing of the examples adheres to the Leipzig Glossing Rules (Comrie, Haspelmath, & Bickel, 1982). Additional abbreviations can be found in the Glossary at the end of this thesis. For pronouns, some "default" glosses are omitted in order to declutter the examples. Unless glossed otherwise, a pronoun is a subject and is realis. These distinctions between pronouns are discussed in Section 3.5.2.

Throughout this thesis, especially in Chapter 2, various Nigerian names will be mentioned (e.g., places and languages). As many of these names have their origin in Yoruba or other Nigerian languages, they exhibit phonemic features like an advanced tongue root (ATR) value or certain pitch distinctions. In Nigeria, these features are usually marked with diacritics, such as the underdot (ọ and ẹ) for [-ATR] and accent marks (e.g., á and à) for pitch. Often, there are no anglicized versions of these names available. In some sources where I found these names (e.g., maps and papers), these diacritics are omitted. In order to prevent typing errors and make writing easier, I will also omit these diacritics in the text and all maps. The only change will be the transliteration of <ṣ> to <sh>, as the postalveolar fricative can be indicated in the English orthography. However, a list of important names written with their diacritics will be given in Appendix B. The names will be presented with a source where they have been indicated with diacritics and, if available, a pronunciation in Abesabesi. These name tables are given to record the full orthographic manifestations of the names. All Abesabesi place names mentioned in the text and the maps will be written in Yoruba spelling without diacritics as this is how they can be found on maps. The dialect names, however, are given in Abesabesi to distinguish them from the town's name (eg. the town Ikaram vs. the dialect Ekiromi). One has to be very careful not to confuse Yoruba spelling and Abesabesi. This is why Abesabesi words are written in italics: *ìjòtù*.

1.3 Coding conventions

Chapter 6 relies heavily on examples of the Extensible Markup Language (XML) or related markup languages. These extracts will be presented in "listings" that are colored in order to facilitate their readability (Listing 1.1). As they are merely extracts, they might not

be valid XML documents per se. Longer examples are attached in Appendix A in order to avoid cluttering the text.

Listing 1.1: XML example

```
<chapter id="1">
        <heading>Phonology</heading>
        <p>Lorem ipsum dolor sit amet, consectetur...</p>
        <section id="1.1">
                <heading>Vowels</heading>
                <p>...</p>
        </section>
</chapter>
```

In-text extracts of markup script (e.g., `<section id="sc1">`), Uniform Resource Locators (URLs) (e.g., `http://www.example.com`), and file-endings (e.g., `.pdf`) will be distinguished from the text by styling them in a typewriter font.

This PhD project features the development of an exemplary grammar exploration tool (GET) (Chapter 7) that can be accessed at `http://abesabesi.cceh.uni-koeln.de/`. Specific code will not be displayed within this thesis as I intend to present the application's overall structure and illustrate the possibilities of digital grammar writing instead of giving an in-depth look at the application's development. However, readers interested in the specific mechanics of the application can access its source code at `https://github.com/JonasLauLau/AbesabesiGrammar`.

This thesis only contains extracts of the encoded Abesabesi Sketch Grammar to illustrate the grammar format and the encoding process. Parts of the Abesabesi Sketch Grammar and related databases (dictionary, text collection, etc.) have been encoded on grammar documents that can also be accessed at `https://github.com/JonasLauLau/AbesabesiGrammar` following the path `resources/data`.

Part I

Abesabesi

Chapter 2

Research Context – Abesabesi

This chapter serves as an introduction to the Abesabesi Sketch Grammar in Chapter 3. It comprises information about the language itself, the field site, the methodology applied in this research, and previous research on Abesabesi. Beginning with basic information on the language, its speakers, and its classification (Section 2.1), this chapter continues with an introduction of the field site and its history (Section 2.2). Section 2.3 then discusses how this research has been undertaken. It follows the process from data collection to the research results by presenting information on fieldwork, data collection, text corpus, language consultants, data processing, and data analysis. Finally, Section 2.4 summarizes findings of previous researchers and other resources on Abesabesi.

2.1 Abesabesi

This section introduces the language Abesabesi, which will be grammatically described in Chapter 3 and later be used as a case study to discuss the development of a DRG. The language with the ISO 639-3 code “ibe” and the glottocode “akpe1248” is known as “Akpes” in literature but will be called “Abesabesi” throughout this thesis. It is spoken by about 7000 speakers in nine settlements in the Akoko area – the northern-most region in Ondo State, Nigeria. Abesabesi has four dialects: Akpes, Ekiromi, Ilueni, and Oshugu. Research for this thesis took place in Ikaram, which is why the grammatical description is based on data of the Ekiromi dialect as spoken in Ikaram. The language is attributed to the Benue-Congo branch of the Niger-Congo phylum, but a closer relationship to subbranches or languages remains disputed. Culturally, most speakers consider themselves part of the Yoruba people. The varieties spoken in the nine settlements are therefore often considered Yoruba dialects, while Standard Yoruba is the lingua franca used for communication within the region and in formal domains within each settlement. Abesabesi is also being replaced more and more in informal and private domains. This has lead to a young generation of 1- to 20-year-olds barely speaking the language and, therefore, to language endangerment.

2.1.1 Naming the language

Abesabesi is known as "Akpes" in literature. However, "Akpes" is the name of one of the four dialects (see Section 2.1.4) and is, therefore, not supported by the speakers as a name for the entire language. If asked to name their language, speakers usually use the dialect's name or the name of their settlement. Generally, a linguistic relation between the different dialects is not always perceived, as many speakers consider them dialects of Yoruba. This is probably rooted in the fact that most speakers consider themselves as Yoruba (see Section 2.1.6). The unity among the nine settlements has been promoted only recently by Dr. Taiwo Agoyi, a native speaker, professional linguist, and the main researcher on the language. In order to chose an acceptable term for the entirety of all four dialects, representatives of the nine communities have agreed on the term *Àbèsàbèsì* to denote the language. This meeting of representatives has become known as the "Àbèsàbèsì Language Development Committee" (see Section 2.1.6). Henceforth, the term has mainly been used by Agoyi. Speakers use it only scarcely but recognize it as a name for the language. It is used in some formal and semi-formal contexts such as in speeches at the local festivals, in language-related meetings and activities, and in names for social media groups dedicated to communicate in the language. *Àbèsàbèsì* is a reduplication of the word *àbès* meaning 'we'. The reduplicated form *àbèsàbès* signifies 'ourselves'.

Throughout this text, "Abesabesi" will be used to refer to the language and the people, as it is a term chosen by representatives of all communities speaking it. As all Yoruba and Abesabesi names and terms will be written without accent marks within this text to avoid spelling errors (see linguistic conventions in Section 1.2), this also affects the term "Abesabesi". A list of all those terms with correct accent marks will be given in the Appendix B. The spelling of "Abesabesi" also differs from the spelling that is used for Abesabesi language data within this text (see Section 3.1.6), where final deleted vowels are not written.

2.1.2 Location and surrounding languages

Abesabesi is spoken in nine settlements in the north of Ondo State, Nigeria. The red dot on Figure 2.1 indicates the location of these settlements within Nigeria and West Africa. They consist of six independent towns (Akunnu, Ase, Gedegede, Ibaram, Ikaram, and Iyani) and three communities within the rather recently founded town Ajowa. Disputes about land and sovereignty has led to a relocation of seven different communities and resulted in the foundation of Ajowa (Agoyi, 2014, p. 4). Of these seven communities, three are Abesabesi-speaking ones: Daja, Eshuku, and Ilodun. Another community of Ajowa, the Efifa, are known to have spoken a dialect of Abesabesi in the past but switched to the local Yoruba variety Owe (Agoyi, 2013b, p. 74). All of the Abesabesi-speaking towns and the different quarters of Ajowa's communities can be seen in Figure 2.2.

The Abesabesi-speaking area is located in the north of Ondo State in an area referred to as "Akoko". This hilly region is composed of four Local Government Areas (LGAs). Ex-

Figure 2.1: Location of the Abesabesi speaking settlements (red dot) in Nigeria[1]

cept Akunnu, which is located in the Akoko North East LGA, all other Abesabesi-speaking communities are within the Akoko North West LGA.

The Akoko region is a highly multilingual area, where eight distinct native languages are spoken next to the dominant lingua franca – Standard Yoruba and its manifold local varieties of the region. Moreover, the region is home to different immigrant communities such as Ebiras, Fulanis, Hausas, and Igbos. Figure 2.3 displays all these languages in a tree structure that has been simplified from (Güldemann, 2018, pp. 353-356). Language names in italics are given if the name used in this thesis deviates from the name used in Güldemann (2018). The genetic affiliation of Abesabesi and alternative approaches to this tree are discussed in Section 2.1.5.

Figure 2.4 displays the different towns and villages of the Akoko region with their respective native language.[1] Yoruba (ISO639-3: yor) is the dominant language of the region. While Standard Yoruba is the lingua franca of the region, different local varieties are native to most of the Akoko towns, especially in the West and the South. The most

[1]All three maps are created by the author. Most of the GPS data was provided by Sophie Salffner. Information about the town-language relations has been taken from several sources – most notably Ohiri-Aniche (1999), Olúmúyìwá and Oshòdì (2012), and Elugbe (1989). The maps attribute one ”native” language to each town. If not possible, towns are marked with ”multiple languages”. Although the decision about which language is ”native” to a town is rather problematic, it is necessary to such a cartographic approach. This means that the maps do not show the lingua francas Standard Yoruba, English, or Nigerian Pidgin English; the languages of various farm settlements and camps; nor the languages of recent immigrants who settled down in the towns.

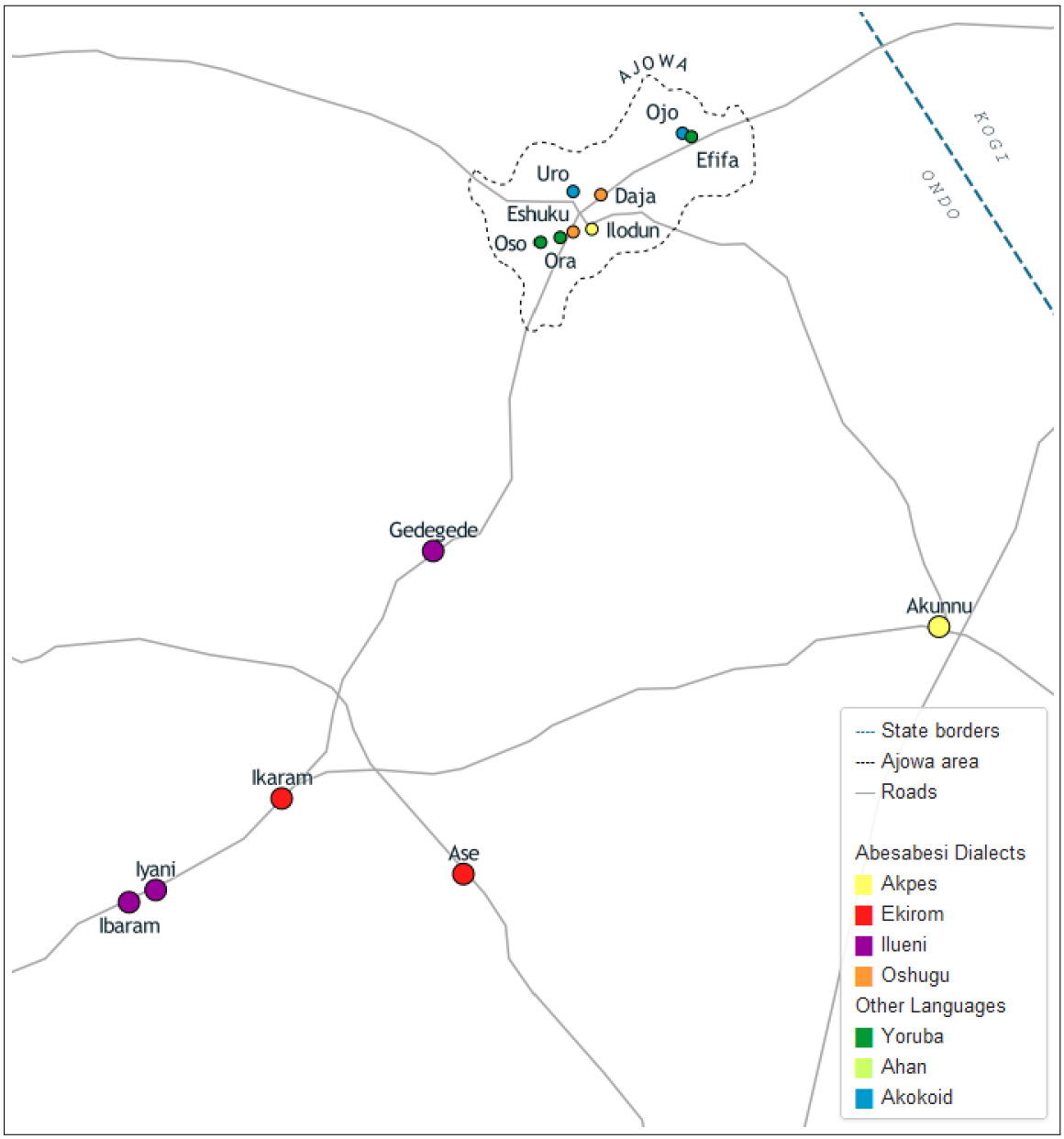


Figure 2.2: The Abesabesi speaking settlements.[1]

notable are Ikare, the largest local town, and Akungba, the home of the local university.

The south of Akoko is home to a couple of Edoid languages of the Osse branch (Elugbe, 1989, p. 24): Ehueun (ISO639-3: ehu) spoken in Epinmi, Ukue (ISO639-3: uku) spoken in Ipe and Igbede Ipe, and Uhami (ISO639-3: uha) spoken in Oyara Ishua, Ishua Oke, Ishua Ile, Shosan Oke, and Shosan Ile. The main area of Edoid languages extends further south and east of Akoko in Edo state.

In the east, the Ukaan language (ISO639-3: kcf) is spoken in Auga, Ishe, Ayanran, and two towns called “Ikakumo”. The language is sometimes also referred to as “AIKA”, an acronym for the five villages, “Auga, Ishe, Kakumo, and Ayanran”. The genetic affiliation of Ukaan remains controversial. It has been attributed to both West Benue-Congo

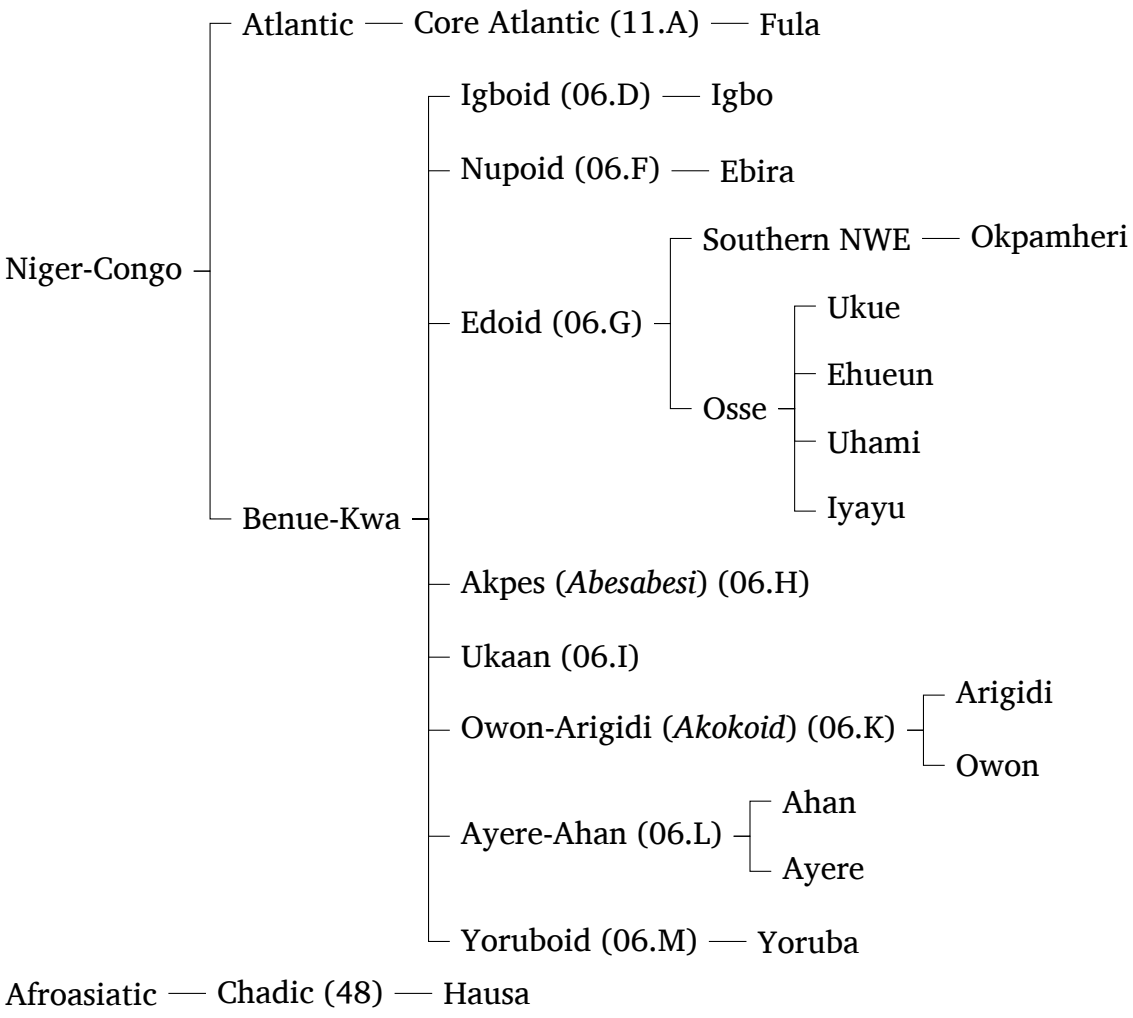


Figure 2.3: Languages spoken in the Akoko region of Ondo State. Identifiers are taken from Güldemann (2018, pp. 353-356)

and East Benue-Congo. Its closest relative based on the lexicon, however, seems to be Abesabesi (Ohiri-Aniche, 1999, p. 83).

In the north, seven towns and three communities within Ajowa (Uro and Ojo) speak Akokoid (ISO639-3: aqg), a language cluster that is part of the Benue-Congo branch. Fadoro (2013) classifies Akokoid into two languages: Arigidi (spoken in Arigidi and Erushu) and Owon (spoken in Eriti, Igashi, two quarters of Ajowa, Oyin, Oke-Agbe, and Iye). The whole cluster was originally grouped together with Ayere (ISO639-3: aye) and Ahan (ISO639-3: ahn) as the "Northern Akoko Cluster" by Hoffmann (1976).

Abesabesi is located in the north east of the Akoko region. It is surrounded by Akokoid in the west, Yoruba in the south, Ukaan in the east, and Ayere in the north. Ikaram, where the data was recorded, is home to the native Abesabesi-speaking population, which is gradually shifting to Yoruba.

In contrast to Abesabesi, which a lot of younger people and immigrants do not speak,

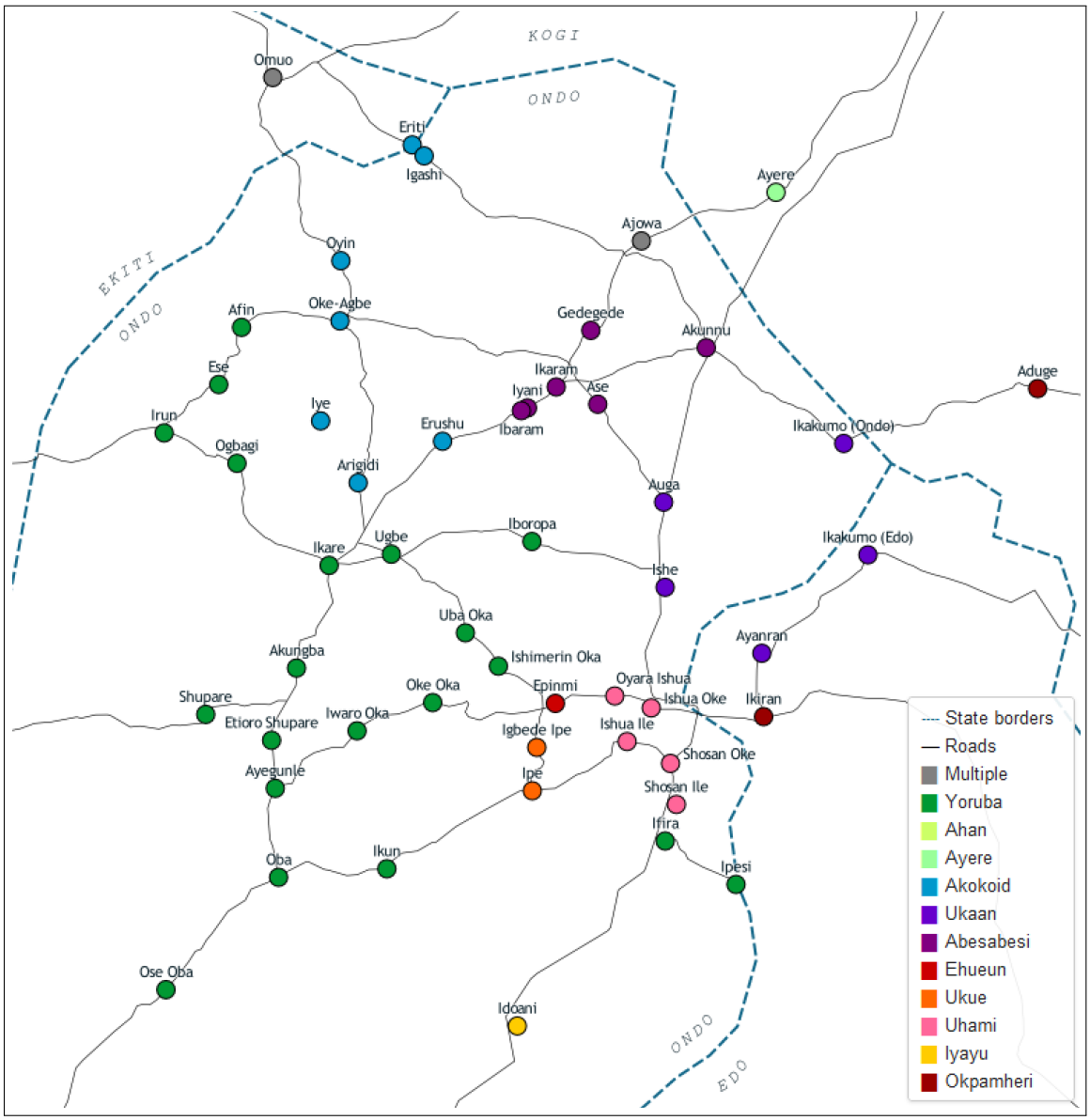


Figure 2.4: Languages of the Akoko towns.[1]

virtually everybody speaks Yoruba. Most people speak Standard Yoruba as it is taught and spoken in schools, and some speak local varieties. Apart from Yoruba, Nigerian English is spoken by people with a higher education. Pidgin English is also common among less educated people, especially in conversations with non-Yoruba speaking people. Other peoples living in Ikaram include Yoruba, Ebira, and Fulani. The Yoruba mainly came as spouses or in higher educated positions such as priests, teachers, or medical personal. Only some of them learned some Abesabesi. The Ebira have lived in Ikaram for a couple of generations. Some Ebira (mainly older people) speak Abesabesi. Nomadic Fulani are present in the whole Akoko region. Some of them have been allowed to settle down in Ikaram in the past. Yoruba, Ebira, and Fulani are generally seen as part of the community of Ikaram, recognize the king (Akala) of Ikaram, and contribute to the town's festivals

with performances, masquerades, or speeches.

2.1.3 Number of speakers

The number of speakers of Abesabesi is not easy to determine. The only existing estimation I could find is 10,000 (M. P. Lewis, Simons, & Fennig, 2018). This estimation is from 1992 and, therefore, quite dated. Agoyi does not present speaker estimations in any of her publications, but gives population census data from the Abesabesi settlements and mentions the difficulties of estimating speaker numbers (Agoyi, 2014, p. 1).

Indeed, there are severe difficulties when it comes to estimating the amount of Abesabesi speakers. On the one hand, only old census data is available and data about languages or ethnicity is not included. On the other hand, the percentage of Abesabesi speakers among the population of the nine settlements can only be estimated. Lastly, quite a substantial part of the Abesabesi live elsewhere in Nigeria or abroad. The percentage of Abesabesi-speakers among these diaspora-Abesabesi, again, can only be an estimate. Nevertheless, I want to attempt an estimate in order to give a more accurate and current view on the speaker number. As an estimation depends on several factors, I will present a transparent description of how I reached the estimated number.

Starting with the population data of all Abesabesi-speaking settlements, the most suitable and current figures seem to be those collected in the scope of the Millennium Village Project Ikaram-Ibaram, a model project to exemplify the Millennium Development Goals of the United Nations (see Section 2.2). Chovwen, Orebiyi, Abdou-Salam, Afere, and Afolayan (2009) give figures for all Abesabesi-speaking settlements except for Akunnu (Table 2.1).

For Akunnu, Agoyi (2008) cites the Nigerian Population Census (NPC) 1991 and gives a population of 2385. As all figures of the NPC 1991 are about double the amount of the more recent Millenium Village figures, I assume a population of 1193 people in Akunnu (half of 2385). As Ajowa is composed of eight quarters, of which three are Abesabesi settlements, I assume a population of 3024 for Daja, Eshuku, and Ilodun (three eighths of 8064). This makes up a total population of 11,393 for all Abesabesi-speaking settlements.

Town name	Population
Ikaram	4982
Ase	72
Ibaram	613
Iyani	514
Gedegede	995
Ajowa	8064

Table 2.1: Population figures by Chovwen et al. (2009)

Speaker percentages are not available – the closest figures are the speaker percentages across age groups in Agoyi (2014, p 4). However, there is no population curve available, which is why speaker percentages can not be derived from those figures. The estimate I purely base on my own impressions and informal interviews with speakers is 50%. This is because most children and young adults under 30 years do not speak Abesabesi, people above 30 years have varying speaker percentages from 10% to 95% depending on the settlement, and most settlements have a small percentage of immigrants that do not speak it either. The estimated amount of Abesabesi speakers in the nine settlements is, therefore, 5697 people.

The Abesabesi population living outside the nine settlements can also only be estimated. From conversations with speakers and personal impressions, I assume the number is rather high, but only very little of the diaspora population speaks Abesabesi, as many of them are born and raised in the diaspora. I am estimating around 1300 more speakers to round up the total amount of speakers to 7000 people. As mentioned before, this estimation of around 7000 speakers is based on a lot of assumptions, personal impressions, and conversations with speakers and should be used with care.

2.1.4 Internal classification

The first to recognize the close relationship of Abesabesi's speech forms was Ibrahim-Arirabiyi (1989), who listed nine "lects" that correspond with the nine settlements. Using the Ibadan-400 wordlist, he collected lexical data from eight of the lects (Iyani is missing). After applying two lexicostatistical methods, he obtained contrary results and states that he was unable to draw a genetic family tree of the lects. The only matching result was that Akunnu and Ilodun had the highest cognate rate. He calls Abesabesi a "language cluster" due to some low cognate rates of under 80% and the lack of mutual intelligibility between some of the lects.

A more recent grouping of the speech forms has been done by Agoyi (2008, p. 6). She proposes that Abesabesi is a language family that encompasses four languages: namely Akpes, Ekiromi, Ilueni, and Oshugu. In later publications, she uses the terms "lect" or "dialect" for these four varieties. As all of these varieties are mutually intelligible (Agoyi, 2001, p. 65) and as they are purely based on differences across location, they will be referred to as "dialects". Their geographical distribution can be seen in Figure 2.2. Agoyi (2008) bases the grouping of dialects on differences in lexicon and phonemics. Her main findings are that each dialect exhibits differing vowel harmony types (Section 3.1.4.4). The four dialects and the settlements they are spoken in can be seen in Figure 2.5.

Akpes is spoken in Akunnu and the Ilodun (or Iludotun) quarter of Ajowa. The latter is the result of an influx of people from Akunnu in 1955 (Agoyi, 2008, p. 2). This is why the quarter used to be called "Akunnu Ajowa".

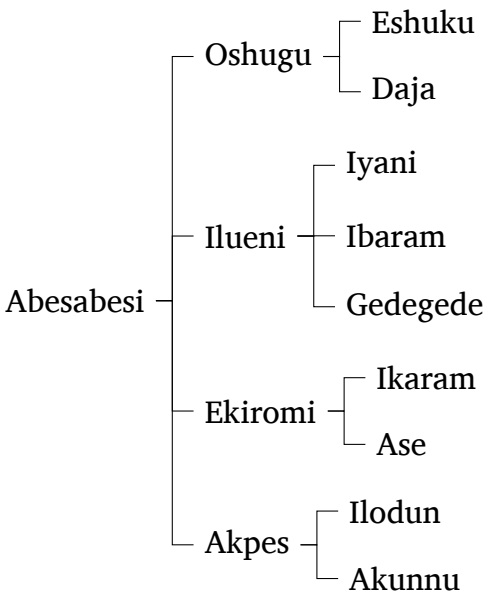


Figure 2.5: Abesabesi dialects and their settlements according to Agoyi (2008)

Ekiromi is spoken in Ikaram and Ase. While Ikaram is a town of almost 5000 inhabitants, Ase is a small farming settlement of less than 100 people about three kilometers east of Ikaram.

Ilueni is spoken in Gedegede, Ibaram, and Iyani. Gedegede is not adjacent to the other two villages but separated from them by Ikaram.

Oshugu is spoken in the quarters Eshuku and Daja of Ajowa.

Efifa, another quarter of Ajowa, used to speak an Abesabesi lect, but its inhabitants switched to the local Yoruba variety Owe. It was still mentioned as an Abesabesi variety in Hansford, Bendor-Samuel, and Stanford (1976, p. 38) and Agoyi (2001, p. 65) mentions that only the elderly still speak Abesabesi. In later publications, she states that the variety is not spoken anymore but remains in placenames and in the local masquerade's singing (Agoyi, 2013b, p. 74). Ibrahim-Arirabiyi (1989, p. 3) already stated that Efifa belongs to the Yoruboid group. He must therefore have worked with consultants unable to speak the Abesabesi variety. It is unclear whether the speech form of Efifa belonged to one of the above dialects or formed a distinct dialect, as it was not included in Agoyi's (2008) internal classification.

This research has been carried out in Ikaram and therefore merely represents the Ekiromi dialect. The gathered corpus (Section 2.3.3), however, contains one wordlist recording of the Ilueni dialect spoken in Iyani.

2.1.5 External classification

The first source to mention Abesabesi was Hansford et al. (1976), where it was marked as "Kwa unclassified". Henceforth, the language has been affiliated with the wider Kwa or Benue-Congo branch of Niger-Congo by all researchers, but a more specific positioning remains controversial. This section merely gives an overview of ways Abesabesi has been positioned within the Benue-Congo branch but does not present a preference, as data on most of these languages is still too scarce to base these genetic relations on sufficient evidence. All attempts to position Abesabesi within Benue-Congo are based on one of three claims: Abesabesi forms an independent branch within (West-)Benue-Congo, Abesabesi is closely related to Ukaan, and Abesabesi is closely related to Ukaan and the Edoid languages. All three claims are discussed in the following paragraphs, illustrating them with one exemplary genetic tree each.

The claim that Abesabesi forms an independent branch of Benue-Congo has been made by Crozier and Blench (1992), Williamson and Blench (2000), and most recently by (Güldemann, 2018), who attempts a more evidence-based approach to the classification of African languages and situates Abesabesi as an independent branch in the wider Benue-Kwa group.

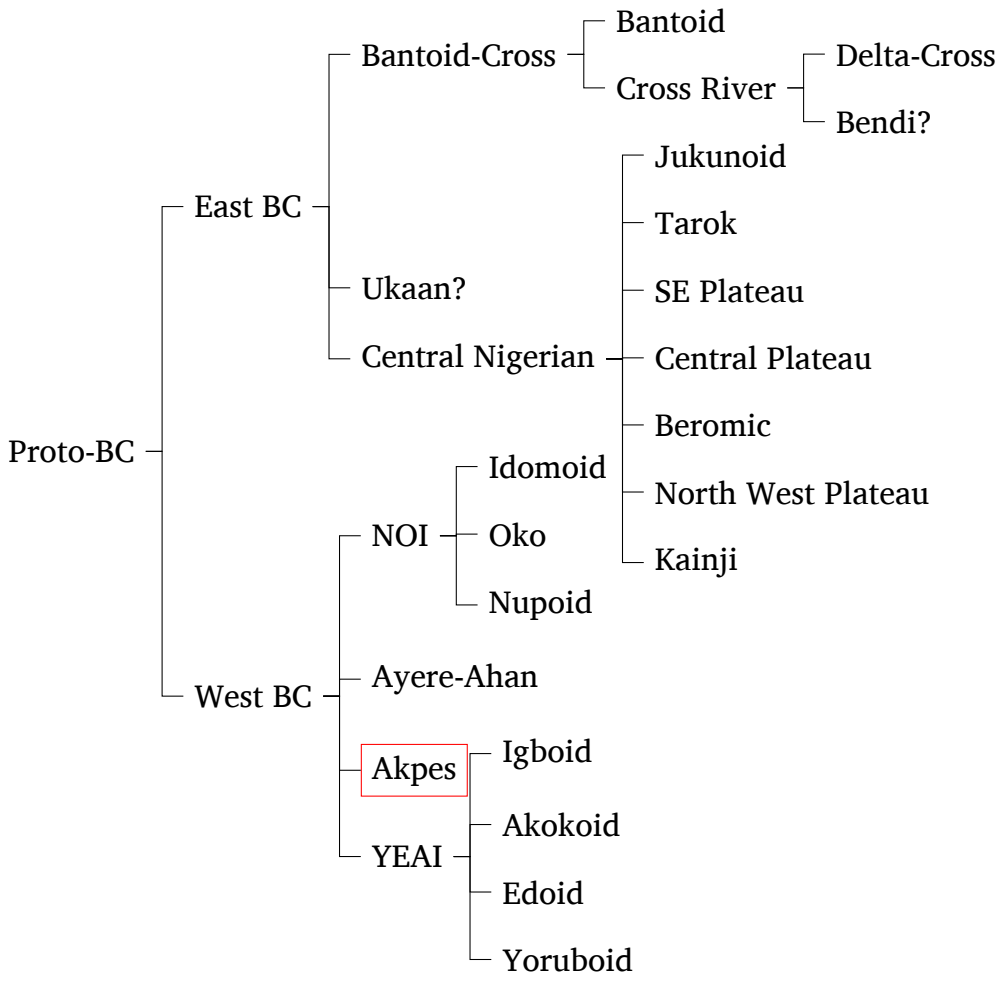


Figure 2.6: Benue-Congo tree according to Williamson and Blench (2000)

There have been two other claims made regarding possible relatives of Abesabesi. While some researchers claim a close relation to Ukaan and position them together in one branch of Benue-Congo, others add Edoid to the branch.

Ukaan seems to be the closest relative of Abesabesi, when only lexical items are taken into account. This was established by Williamson (1989, pp.266-267), who based her assumptions on data provided by Ibrahim-Arirabiyi (1989) (Güldemann, 2018, p. 165). Later, Ohiri-Aniche (1999, p. 84) conducted a lexicostatistical survey comparing Abesabesi, Akokoid, Ukaan, Edoid, Nupe, Standard Yoruba, and the local Yoruba variety Oka. She calculates a cognate score of 45% between Abesasbesi and Ukaan and even mentions some similarities to the distant Jukunoid Family. Salffner and Sands (2012, p. 4), however, point out the difficulties of comparing the under-described languages, Ukaan and Abesabesi, especially if classifications are based on a single speaker of only one dialect.

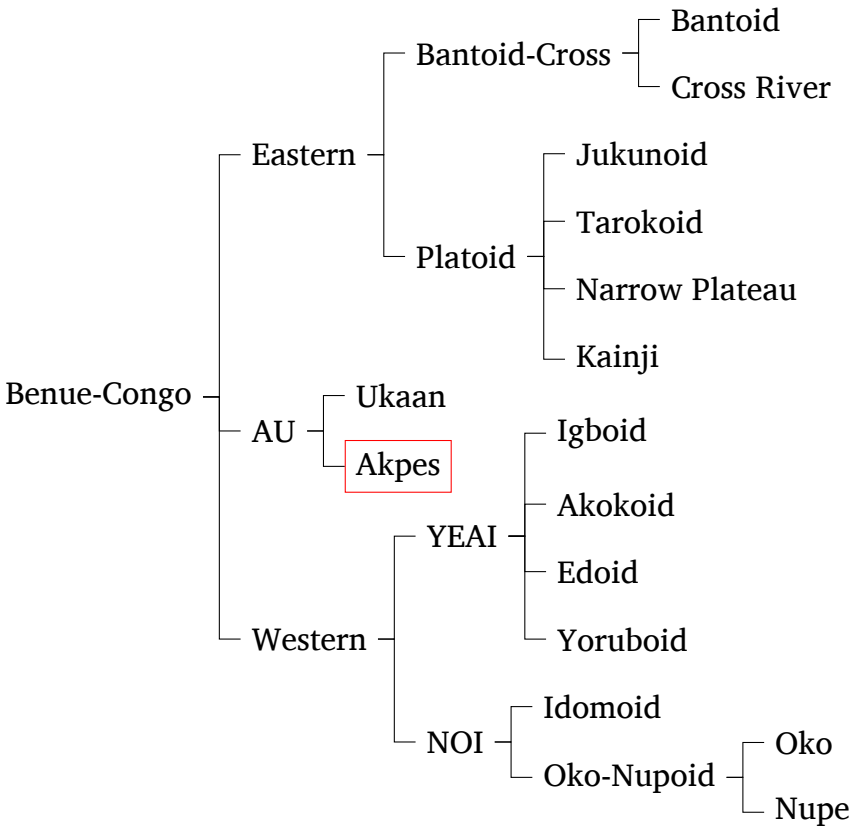


Figure 2.7: Benue-Congo tree according to Ohiri-Aniche (1999)

Agoyi (1998) claims that Abesabesi is affiliated with the Edoid languages (cited in Agoyi (2008, p. iv)). She compares the initial vowel alternations to mark the plural – remnants of a noun class system – in Abesabesi, Ukaan, and Edoid and concludes that all three lineages form one language family. This claim is repeated in Agoyi (2001). Elugbe (2012) names this proposed language family "Akedoid" and treats Abesabesi and Ukaan as earlier branchings of the tree that later becomes Edoid (see Figure 2.8). As evidence, he lists cognate rates, regular sound changes in four items, phonemic innovations of the proposed family, and a similar gerund morpheme with a matching tone pattern. Apart

from the scarcity of the data he presents, his single example for the Abesabesi gerund is rather confusing, as it depicts a gerund pattern that is not known to me nor does it appear in any other publication I am aware of. Agoyi (2013a) backs his claims, repeats her earlier hypothesis about vowel alternations, and presents more accurate gerund data that displays the similarity of the Abesabesi gerund to the Edoid gerund constructions. Her Abesabesi tone analysis, however, is rather confusing and seems to contradict her claims. Güldemann (2018, p. 165) assesses these claims and concludes that they are based on "restricted and inconclusive evidence".

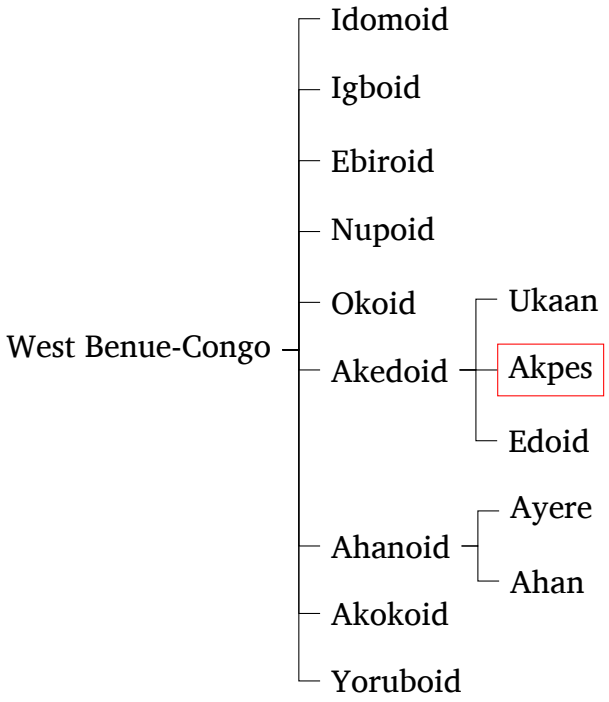


Figure 2.8: West Benue-Congo tree according to Elugbe (2012)

Apart from the fact that Abesabesi has high cognate rates with Ukaan and Edoid, all of the presented proposals contain weak hypotheses that are based on restricted data. More structural methodology and analyses of more dialects are needed to specify the positioning of Abesabesi within Benue-Congo. Until then, the genetic affiliation of Abesabesi remains unclear.

2.1.6 Sociolinguistics

This section will be a brief outline of topics that are concerned with the interaction between Abesabesi and the society it's spoken in – from language acquisition and language use to attitudes and endangerment to literacy. Apart from comments in some publications by Agoyi (2008, 2014), the only data on sociolinguistic topics regarding Abesabesi is Oyetade (2004). This survey was carried out as a questionnaire with a sample size of 234 participants from four of the settlements (Akunnu, Daja, Gedegede, and Ikaram) enhanced by observations in several locations to verify language choice. Although it is

outdated and the sociolinguistic situation can change rapidly, this will be the main source of this section, as it is the only empirical source available.

Acquisition and proficiency Oyetade's (2004, p. 535) findings are that Yoruba and Abesabesi are acquired simultaneously at home and in the neighborhood. While the self-rated overall proficiency (speaking and understanding) in his survey was better in Yoruba than it was in Abesabesi, rates showed differences depending on several parameters. Yoruba proficiency rates are higher among better educated participants. Oyetade sees the Yoruba-dominated education system as a reason for this. As Yoruba is the medium of instruction in primary schools and, at the same time, a subject in primary and secondary schools, proficiency is expected to continue increasing. Abesabesi, on the other hand, is not taught in school, which is reflected in the proficiency rates as no correlation with education could be found. Older people exhibited better proficiency rates in Abesabesi and Yoruba. Different genders did no show any differences in proficiency. Agoyi (2014, p. 4) lists the percentage of each settlement's Abesabesi-speaking population according to age group. The low rates between ages 1-20 (0-11%) reflect Oyetade's findings. Speaker rates increase with age to 57-95% among all people above the age of 50. Distinguishably lower speaker rates in all age groups can be found in the settlements located in Ajowa (Daja, Eshuku, and Ilodun).

Language use Located in one of the six Yoruba states (Lagos, Ekiti, Ogun, Ondo, Osun, and Oyo), the Abesabesi have been "integrated into the Yoruba soocio-cultural [*sic*] milieu" (Oyetade, 2004, p. 532). Almost all of the consultants I have worked with, for example, identified as Yoruba and many consider Abesabesi to be a dialect of Yoruba. The locally produced history booklet of Ikaram (Abimboye, 2008, p 15) reflects this view. The use of both languages results in a daily bilingualism, each language being stronger in certain domains. English, as the official language of Nigeria, competes with Yoruba in more formal domains. In Oyetade (2004, p. 537), language choice has been shown to be split mainly between Yoruba and Abesabesi. The majority of children claim to use Yoruba amongst family, even with grandparents. In school, the majority claimed to speak English with teachers, a mix of Yoruba and English with classmates, and Yoruba on class breaks. The majority of parents stated they use predominantly Yoruba with children – 50% exclusively Yoruba. Also with their spouses, Yoruba is the preferred language. Oyetade assumes frequent exogamy as a reason for this. Marriages with Yoruba might not be perceived as exogamy, since Abesabesis often identify as Yoruba. This might be a reason for the frequency of exogamy. According to Oyetade, parents reported choosing to speak Abesabesi or Abesabesi and Yoruba in conversations with relations, neighbors, other speakers of the same dialect, in community meetings, and when discussing family matters. Yoruba, however, is not weak in those domains. Community meetings, for example, have the lowest percentage of people stating to speak only Yoruba with 25%. In all other domains, more people claim to speak only Yoruba. Oyetade concludes that Yoruba is much stronger in these informal domains, as one would expect, and sees a

trend that might become even more fatal for the language. Considering the survey is more than 15 years old, Yoruba might be even more dominant in all domains at present. In all formal domains, Yoruba is already dominant. In church and mosque services and social institutions, only Yoruba is spoken (Oyetade, 2004, p. 541). Oyetade states that Yoruba has long been the language of trade, and as markets are very heterogeneous in terms of language and country of origin, Yoruba is the default for all market transactions. Only if local merchants know their customer, and both know that the other speaks Abesabesi, Abesabesi is spoken. Oyetade's results differ from Agoyi's (2008, pp. 1-6) comments on language use in two points. Agoyi states that Abesabesi is dominant in market interactions and conversations within the family.

Attitudes and identity Oyetade (2004, pp.541-542) included statements in his participant survey that dealt with attitudes toward Abesabesi and Yoruba as well as language identity. He asked the participants to indicate whether they agreed or disagreed with the statements. Although Oyetade did not write the exact statements, they were paraphrased along the lines of "My Abesabesi variant is an index of my identity" (1), "Yoruba is enough for my identity" (2), "Children should be encouraged to learn their Abesabesi variant" (3), "Yoruba is a thread to Abesabesi" (4), "Yoruba will replace Abesabesi if no measurements are taken" (5), "Abesabesi speakers are inferior to other speakers of Yoruba" (6), and "I am ashamed to speak Yoruba in mixed gatherings" (7). In response to the first statement, a large percentage of participants in each settlement agreed that they consider their own variety of Abesabesi to be a part of their identity. However, a large percentage of participants also agreed with the second statement. The only notable exception was Ikaram, where 79.2% of participants from this settlement disagreed that Yoruba is how they describe their identity. A large percentage of the participants also agreed with Statements 3, 4, and 5, which expressed how Yoruba is a thread to Abesabesi, how Yoruba will replace Abesabesi in the future if no measurements are taken, and how it is important to encourage children to speak Abesabesi. This shows that people are aware of Abesabesi's looming death and find maintenance measurements necessary in order to prevent it. On the other hand, the survey also shows unfavorable language attitudes. The majority of participants thought Abesabesi to be inferior to Yoruba (6) and were also ashamed of speaking Abesabesi in mixed gatherings (7).

Literacy As Yoruba and English are taught in school, literacy in the Abesabesi-speaking settlements is restricted to those two languages. Although literature in Abesabesi does not exist, written Abesabesi exists in social media groups dedicated to the language. While working with language consultants, the Yoruba alphabet was preferred for spelling Abesabesi words, as Yoruba phonology is much more similar to Abesabesi than English phonology. In social media conversations and text messages, I observed an anglicized version of the Yoruba alphabet when writing Abesabesi. This is likely due to the lack of Yoruba keyboards on mobile phones and computers. The anglicized version omits the dots under <ẹ> and <ọ> and replaces <ṣ> with the English grapheme <sh>. Tonal

markings are very rarely employed. This is also a common way to write Yoruba in social media and text messages.

Agoyi (2013c) proposes an orthography for all Abesabesi dialects. The orthography is based on the Yoruba alphabet, which is familiar to all Abesabesi speakers. Although she claims the orthography can be applied to other dialects than Ekiromi, the paper exclusively uses Ekiromi examples. Besides a phoneme inventory with corresponding graphemes, she explains how the orthography handles the following phonological processes and suprasegmentals: Nasalization, long vowels, tones, and vowel deletion. A few phonological phenomena and their orthographic representations, however, are omitted. These include assimilation, vowel harmony, epenthetic vowels, contour tones, downstep, and grammatical tone. The proposed orthography is mainly used by Agoyi in her publications.

For this thesis and the entire corpus, I have developed a practical phonemic orthography that is based on the International Phonetic Alphabet (IPA) (Section 3.1.6). It could, however, be adapted to the Yoruba alphabet and is thus similar to Agoyi's proposed orthography.

Vitality and endangerment assessment Language vitality or endangerment can be assessed through different scales or questionnaires. An often used scale has been the Graded Intergenerational Disruption Scale (GIDS) (Fishman, 1991, pp. 89-109), which was later expanded to the Extended Graded Intergenerational Disruption Scale (EGIDS) by M. P. Lewis and Simons (2010), making it comparable with the scale introduced in UNESCO (2003).

Existing endangerment assessments of Abesabesi are rather dated and differ between "safe" and "severely endangered". M. P. Lewis et al. (2018) assign the EGIDS level 6a (Vigorous), which is the last level indicating a safe language vitality. Oyetade (2004, p. 545) classifies Abesabesi according to the criteria of Wurm (1998) as a potentially endangered language that is "socially and economically disadvantaged, under heavy pressure from a larger language, and beginning to lose child speakers". Agoyi labels Abesabesi "endangered" or "highly endangered" in her publications (Agoyi, 2012b, 2013b, 2014, 2015a) without referencing a scale. She justifies the label due to the discontinuation of intergenerational transmission.

To give a current view on Abesabesi's endangerment, the EGIDS questionnaire will be used. It comprises five key questions to determine the degree of endangerment. All answers to the questionnaire will be explained and justified in the following paragraphs. As answers to the questions lead to other questions and finally lead to a final result, only three questions are addressed here before resulting in an endangerment label.

Key Question 1: What is the current identity function of the language? As Abesabesi is not used as a vehicular language to facilitate communication between speakers of different L1s, and L1 speakers of Abesabesi still exist, the answer to this question is "home". This answer takes us directly to Key Question 3.

Key Question 3: Are all parents transmitting the language to their children? The answer to this question is "no". As seen earlier in Section 2.1.6, a substantial amount of parents exclusively use Yoruba with their children. The low speaker rates for children given by Agoyi (2014, p 4) reflect that. This answer takes us directly to Key Question 5.

Key Question 5: What is the youngest generation of proficient speakers? This question is harder to answer. Oyetade (2004, p. 536) does not display his proficiency figures for the different ages, but only mentions a positive correlation between age and proficiency rate. Agoyi (2014, p. 4) gives speaker rates of 0% within the age range of 1-20 years in the Abesabesi-speaking quarters of Ajowa (Daja, Eshuku, and Ilodun) and rates of 0-11% in all other Abesabesi-speaking settlements. As I was not able to find a single consultant younger than 24[2], I assume that the speaker rates have dropped since 2014. This is why I chose the answer "parents", for which is stated that "[t]he youngest proficient speakers of the language are the adults of child-bearing age".

This results in the EGIDS Level 7 (Shifting). This level corresponds with Grade 3 (Definitely Endangered) on the UNESCO scale, which has been evaluated separately in table 2.2 (Factor 1). The two results assessed here display a more severe endangerment status as those assessed in M. P. Lewis et al. (2018) and Oyetade (2004).

2.1.7 History

Apart from recent events and migrations within the Akoko region, the history of the Abesabesi-speaking towns has exclusively been passed on orally. Agoyi (2013b) attempts to summarize these various stories about the origins of the different communities. As each settlement seems to be composed of different quarters or sub-communities with their own history, this attempt is rather hard to accomplish. The Abesabesi-speaking settlements seem to be conglomerates of various communities that migrated from different places and joined each other at the present settlements or even earlier. Which migration lineage brought the language that unites all settlements is unclear. One migration lineage, however, can be found in several of the settlements' legends. It involves an origin in Ile-Ife (present day Osun State) and a group of related founding fathers that left Ile-Ife in search of greener pastures. While tales of different stops along the way to the current settlement locations have been passed down, it is believed that these founding fathers eventually arrived at and founded each of their respective settlements/quarters which coincide with their present day locations. As Ile-Ife is seen as the cradle of the Yoruba and constitutes the setting of the Yoruba creation myth, parts of the legend might have been introduced by the Yoruba.

In the course of colonization, the region was traversed by British troops. According to Abimboye (2008, p. 19) and Agoyi (2014, p. 6), the Royal Niger Company (RNC)

[2]The youngest consultant only happened to speak Abesabesi because he grew up with his grandparents. All other consultants were above the age of 30.

	Factor	Grade	Degree of endangerment
1	Intergenerational language transmission	3	Definitely Endangered
2	Absolute number of speakers	7,000	
3	Proportion of speakers within the total population	3	Definitely Endangered
4	Shift in domains of language use	3	Dwindling Domains
5	Response to new domains and media	1	Minimal
6	Availability of materials for language education and literacy	1	A practical orthography is known to the community and some material is being written.
7	Governmental & institutional language attitudes and policies including official status & use	?	
9	Community members' attitudes towards their own language	3	Many members support language maintenance; others are indifferent or may even support language loss.
10	Type and quality of documentation	2	Fragmentary

Table 2.2: Vitality assessment of Abesabesi (the author's assessment using the UNESCO scale (UNESCO, 2003))

established barracks in Ikaram, and Lieutenant Francis Sillifant settled down there as well. His former house and his tomb can still be seen.

A rather recent event is the Millenium Village Project Ikaram-Ibaram, started in the course of the African Millennium Villages Initiative. This project, funded by the Earth Institute at Columbia University, amongst others, was planned out in order to test the feasibility of the UN Millenium Development Goals. A mid-term report (Boyd, Asiabuka, Medupin, & Osunsanya, 2009) and a paper assessing the water and sanitation intervention of said project (Chovwen et al., 2009) were accessible to me and provide useful data about geography, population, and infrastructure.

2.1.8 Typological sketch

This section aims at a brief overview of the structure of Abesabesi. It also presents several salient features of the language.

Abesabesi is a tonal language with three tonal registers. All three tones are used to distinguish lexemes, while the high tone can only rarely be found on lexemes. It does, however, function as a grammatical tone marking the mood of a clause, locative phrases, possession, and relative clauses. Automatic and non-automatic downstep both exist in Abesabesi. The vowel inventory consists of seven oral vowels and five nasal vowels, and the consonant inventory includes labio-velar plosives and several labialized consonants. Frequent phonological processes are vowel deletion, assimilation, and vowel harmony. The latter can be categorized into a prefix and a suffix harmony that take different forms in the four dialects. Abesabesi's syllable has the structure N or (C)V(V)(C). Closed syllables, however, only appear at the end of a word. Nouns start with a vowel and have the minimal structure V.CV, and verbs have the minimal structure CV. Verbs and nouns can both take various inflectional and derivational affixes. Abesabesi could thus be counted as an agglutinative language.

Nouns are categorized as "human" and "non-human". While human nouns switch their initial vowel for plural marking, non-human nouns do not mark plurality. Determiners and property nouns are also marked as plural by switching the initial vowel. Nouns can take possessive suffixes and a goal suffix. All nominal modifiers (other nouns or pronouns, determiners, numerals, and property nouns) follow the noun in a noun phrase. Spatial relations are expressed through function nouns that are located before the noun.

Verbs can take subject pronoun prefixes, TAM prefixes, and object pronoun suffixes. Moreover, they can receive a pluractionalis extension or be nominalized by either a prefix (simple nominalization) or a circumfix (gerund). The bound subject pronouns come in three sets: realis pronouns, irrealis pronouns, and habitual pronouns.

Besides the three bound pronoun paradigms, there are the following personal pronoun sets: independent pronouns, logophoric pronouns, dative pronouns, independent possessive pronouns, and emphatic pronouns. All personal pronouns distinguish three persons, singular from plural, and in the third person, human from non-human.

The basic order of the sentence is SVO. Ditransitive verbs are followed first by the indirect object and then by the direct object. A closed class of auxiliary verbs can precede the main verb or follow it in order to add context to the verbal event or to add a participant.

Abesabesi distinguishes the three moods – declarative, interrogative, and imperative – which are differentiated by grammatical tones. Negation is expressed through the absence of a grammatical tone and a clause-final particle.

2.2 The field site

Fieldwork for this project was carried out in Ikaram, a town of around 5000 inhabitants. This means I was based in Ikaram and recordings took place in Ikaram with speakers from Ikaram. It is located at 7°36'38.2896"N, 5°51'49.0752"E in the Akoko North West LGA, the northernmost LGA of Ondo State in Nigeria. It stretches along a road that connects Owo in Ondo State with Kabba in Kogi State, which can be used to drive between Lagos and Abuja. As in other parts of the Akoko area, Ikaram is surrounded by hills, forest, and farm land. Boyd et al. (2009, p. 47) classify the area as "derived guinea savanna" with a rainy season from March/April to October/November. The economy is characterized by farming, and virtually everyone cultivates farmland – some for a living and others for their own subsistence. Cash crops are yam, cassava (mainly in the form of *gari*, the fermented and roasted cassava flour), and corn. Crops that are mainly used for self-subsistence are melons, guinea corn, tomatoes, pepper, beans, and various leafy vegetables. Livestock include goats, sheep, turkeys, and chickens that are held within the village. Cattle is held by Fulani tribes outside the village. Ikaram has two markets that take turns being open and operate every five days.

Ikaram has a relatively well-developed infrastructure. It is situated on a main road connecting major cities, has daily public transport to regional centers, a health center, several primary and secondary schools, and access to the national electricity grid. However, infrastructure problems that are common to the whole nation, such as continual power failures, poor conditions of streets, and the lack of safety on roads due to armed robbers and kidnapping, also affect Ikaram.

The town is composed of five quarters: Ayeteju, Itabalogun, Iyoke, Iyotu, and Oyagi. Two of those quarters have a different name in the Ekiromi dialect of Abesabesi: Ayeteju is called *ɔ̀ʃùgù* and Itabalogun is called *ìtùjò*. Agoyi (2013b, p. 72) mentions that each of these quarters used to be independent with their own king until 1885, when the Akala Akombo Momodu of Iyotu joined them together and started to rule over the whole of Ikaram. *Akala* is the title given to the king of Ikaram. Abimboye (2008, pp. 7-9), however, claims the founding father of Iyotu already arrived with a crown and thus ruled over the whole of Ikaram. He presents a line of succession starting with the son of the founding father and ending with the ruling king at the date of publication. It includes 12 kings. The last king has died and is followed by the presently ruling king, the Akala Andrew

Momodu.

In terms of religion, the people of Ikaram are Muslim or belong to one of the various Christian denominations that are popular in Nigeria. While Islam reached the Akoko area through the Nupe invasion, Christianity entered Nigeria's Southwest through European missionaries and expanded through Yoruba missionaries to the Akoko area. I had no access to figures, but I am estimating that about a half of the population is Muslim and the other half Christian. Both religions are rather common in the Yoruba dominated area of Nigeria. There are elements of the traditional beliefs, which are still practiced today, but usually only in addition to the Christian or Muslim religion. These traditional beliefs are based on the Yoruba religion and incorporate local deities and customs. The local deity of Ikaram is called *òtòlòm* and is attributed to a pond within the town area. It is worshiped at a dedicated shrine, where a chief priest is in charge of ceremonies.

Customs specific to the place are a local masquerade called *abug* and a festival *abog*. The festival had been a celebration of mature girls of marriageable age, but it is not celebrated anymore. Dances associated to the festival, however, are still performed at the New Yam Festival (*imùnaǹ* in Ekiromi).

2.3 Methodology

The Abesabesi related findings within this thesis stem from data collected within the ELDP funded documentation project "Documenting Abesabesi". The project included two fieldtrips of three months each in 2019 and the compilation of an audio-visual corpus. Apart from the corpus, one session recorded prior to this documentation project by Sophie Salffner has been annotated by me and added to my text corpus. This is why I am using my internal identifier ibe001 for it instead of using Salffner's identifiers. The file itself can be found in Salffner's (2010) deposit at `https://elar.soas.ac.uk/Record/MPI1063487`. Other language data has been taken from examples in Agoyi's various publications. As transcriptions are very inconsistent and primary data is not available, this data is only used to compare it to the data of my own corpus.

2.3.1 Fieldwork

During the fieldtrips, I was based in Ikaram. Due to an unstable safety situation and an increased number of kidnapping cases in the wider region at that time, I was not able to take part in the town life and interact with speakers as much as I intended. Most of my time was spent at the palace of the king of Ikaram, where I lived and held the recording sessions and other meetings with the consultants. This recording location decreased the naturalness of the data, as the language was only recorded in a staged setting. On the other hand, the recording conditions were stable and could be optimized to achieve better audio and video quality. As not everyone in the palace spoke Abesabesi and my interaction with people in the town was limited, my personal proficiency remained on a

basic level.

2.3.2 Data collection

The documentation project was intended to follow the framework of Himmelmann (1998) and Woodbury (2003), but deviated from it in one substatial factor: Communicative events had a low level of naturalness. Out of the three types of communicative events Himmelmann (1998) deems amenable for language documentation, elicited and staged communicative events make up the vast majority of the corpus. On only very few occasions, I was able to record ceremonies and speeches in a public setting and therefore was able to observe spontaneous communicative events. The usual setup consisted of a consultant or two sitting at a table and me sitting on the other side of the table facing them. The camera was used in all staged events and in some elicitation sessions. It was set up next to me, but faced the consultants. Elicitations required an interaction between the consultant and myself, while staged monologues were directed towards the camera. For staged dialogs, the two consultants usually turned toward each other to be able to interact with each other while being filmed from the other side of the table. While all sessions involving two speakers were recorded with a stereo microphone, all sessions with only one speaker were recorded with a mono microphone. In order to translate and transcribe recordings, I used a method inspired by the BOLD approach (Reiman, 2010). I replayed single segments of the recording to one consultant. After replaying, I asked the consultant to slowly repeat the segment and translate it. If parts were unclear to me, we discussed it, or I conducted some contrastive elicitation to distinguish a specific morpheme and its use. Most of these sessions were done with one specific consultant that particularly liked this kind of work and was skilled at it. I did not film these sessions but recorded the audio and archived them as "help sessions". All interaction with the consultants took place in Nigerian English. Yoruba also occurred frequently within the sessions, as it is the lingua franca of the region and influences the Nigerian English of the region heavily. Yoruba was also used for back translations in order to verify semantics or grammatical meaning, as it is much closer to Abesabesi than English. My own proficiency of Yoruba is basic.

2.3.3 Corpus

Most of the data collected during the fieldtrips was archived in the Endangered Languages Archive (ELAR). Apart from the recorded consent sessions containing personal data, everything is openly accessible at `http://hdl.handle.net/2196/00-0000-0000-0012-6E05-F`. Recordings of the examples used in this thesis can be found in the ELAR deposit. In order to find the corresponding files, consult the file-naming conventions in Section 1.2. The corpus is composed of 360 recording sessions and contains 57 hours of video recordings and 75 hours of audio recordings. The sessions were divided into the following categories:

- Consent (20 sessions): besides the general informed consent session, I recorded a consent session with every consultant.
- Elicitations (65 sessions): elicitations include lexical elicitations (wordlists) and grammatical elicitations in various forms. Translation tasks make up the vast majority of the elicitations (especially for wordlists and paradigms), but other methods, such as completion tasks and grammaticality judgments were also used. Some questionnaires were composed by myself, other standard questionnaires were taken from literature.
- Help-sessions (72 sessions): this method involved the slow repetition of speech and translation in order to facilitate annotation.
- Monologues (134 sessions): monologues involved one consultant speaking to the camera. The topics were either chosen by the consultant or suggested by me. In some cases the topic was given a few days in advance for the consultant to prepare the monologue.
  - Personal narratives: consultants speaking about themselves, their families, or important events in their lives.
  - Traditional stories: Consultants narrating traditional stories, mainly Abesabesi translations of traditional Yoruba stories.
  - Procedurals: consultants describing food production, processing and use, material production and use, or traditions and events.
  - Proverbs and explanation: consultants reciting proverbs and explaining their meanings.
- Stimuli-reactions (22 sessions): different stimuli have been used to evoke quasi-natural speech.
  - Story Builder Cards[3]: these cards were designed to facilitate creative storytelling. They depict people and actions and are supposed to help consultants invent a story. I used either the action cards to let the consultants describe the actions one by one or used one character card and three action cards to let the consultant invent a story.
  - Pear story: a video clip with sounds but without spoken language was played before the consultants were asked to retell the story.
  - Frog story: a story told in several pictures was shown to the consultants, before they were asked to retell the story.
  - The Northwind and the Sun: an animatated video clip with voice actors telling a story in English was shown to the consultants before they were asked to retell the story.

[3]retrieved 2021-01-25, from `http://www.story-builder.ca/`

- Dialogs (26 sessions): dialogs involved two people speaking about a topic to each other. The topic was either chosen by the consultants or suggested by me.
- Other (21 sessions): these include ceremonies, performances, songs, and an interview.

For empirical research, I used all the transcribed texts I imported to the program Fieldwork Language Explorer (FLEx)[4]. This text corpus contains 11,900 segments, 2222 types, and 53,797 tokens. FLEx was also used to collect a dictionary of 1200 items.

2.3.4 Language consultants

Throughout my fieldwork, I worked with 20 speakers (see Table 2.3). Apart from one consultant who lives in Arigidi, all speakers live in different quarters of Ikaram. Of the 20 speakers, 6 were female and 14 male. However, one of the women was closest to being my main consultant as she produced the most non-elicited recordings of all the speakers. As each speaker had preferences for specific kinds of work, the genres exhibit a bias in terms of gender. For example, elicitation recordings were mainly done by men, as all women preferred to record narratives. As for age differences, the corpus includes speakers between the ages of 24 and 74. The older group of speakers were nine people between 59 and 74 years old. These speakers have provided most of the non-elicited data and all of the traditional stories. The second age group, 10 speakers between the ages of 30 and 44, could be counted as the last generation of speakers that have learned Abesabesi as their first language. Younger people in Ikaram might understand Abesabesi but usually do not speak it. One exception was my youngest speaker of 24, who grew up with his grandparents and spoke Abesabesi with them. In terms of social status and education, there is a strong bias in my corpus. Almost all speakers of the older age group have a higher degree, had regular employment before retirement, and hold a religious or political chieftaincy title. My three main consultants, who provided about half of all data, are retired teachers. They were most comfortable doing monologues and telling traditional stories. Within the second age group, almost all people have no higher degree and work either as farmers or are in low paying occasional employment. One exception is the consultant who did most of the help sessions. He has a university degree and works as an architect. All consultants stated that Abesabesi is their first language and Standard Yoruba their second. Basic proficiency in (Nigerian) English has been acquired by everyone in school. The speakers with a higher education have a higher proficiency in English. This was important for elicitation, which was conducted in English. Several consultants speak a fourth language such as Ebira, Hausa, Nigerian Pidgin English, or Arabic and some speak a local variety of Yoruba.

[4]retrieved 2020-10-12, from `https://software.sil.org/fieldworks/`

Code	Name	Age (in 2019)	Sex
04	Janet Akeju	63	female
05	Adeoye Blessing Oluwasheyi	31	male
06	Abiodun Afolabi	67	male
07	Steven Ayodele Awolami	72	male
08	Bashiru Naibi Aminu	64	male
09	Arohunmolase Haruna	44	male
10	Caroline Elegbeleye	64	female
11	Alice Elegbeleye	59	female
12	Olubunmi Victor Daudu	33	male
13	Medinat Folashade Oludade	43	female
14	Laisi Rafiu	36	male
15	Adeolu Mercy Funke	44	female
16	Alamo Idris Alooma	38	male
17	Shittu Samuel Aderemi	24	male
18	Ajatta Adebowale Ilesanmi	30	male
19	Friday Jimo Ibrahim	40	male
20	Weehab Aminu Oshono	74	male
21	Olaniran Samuel Jimo Eleperi	61	male
22	Victoria Aderinsọla Ayetoro	64	female
23	Kayode Seth Banfe	32	male

Table 2.3: Overview language consultants

2.3.5 Data processing and Annotation

Metadata for all recording sessions was captured using the IMDI format and can be accessed through the archive's website. The primary files remain unprocessed – only audio files have occasionally been cut into two or more subfiles to facilitate annotation. Annotations were made on 104 of the 360 files. These include transcriptions, translations, and grammatical annotations. In order to produce time-aligned annotations for the recordings, the annotation software ELAN[5] was used. Segmentation, transcription, and translation was completed in ELAN. Grammatical annotation was added in FLEx and the result was integrated back into ELAN in order to have time-aligned annotations with the respective recording. Transcription and translation of elicitations were done by me except for a few where a consultant helped me. All non-elicited texts were transcribed and translated in recorded help-sessions with a consultant. All transcriptions used in this thesis have been done using a practical phonemic orthography, which will be presented in Section 3.1.6. The first lexical items and simple sentences were collected in elicitations. Additional short texts were recorded using stimuli such as the pear story or the Story Builder Cards. These texts were transcribed and analyzed in order to find grammatical categories. Paradigms of these categories were, in turn, collected in following elicitation sessions. These steps were repeated with a growing complexity and naturalness of texts. Elicitations on word order, phrase constituency, and pragmatic use complemented this work-flow. The growing corpus of transcribed texts, which represented a higher number of speakers, helped facilitate grammatical description.

2.3.6 Analysis

The grammatical description of Abesabesi presented in Chapter 3 is based on the data of the text corpus produced in the documentation project (Section 2.3.3). For the phonological analysis, I have transcribed recordings, searched for minimal pairs, and investigated the phonological contexts of particular phones or phonological features. The analysis has been enhanced through the software programs Praat[6] (spectral, formant, and pitch analysis) and Phonology Assistant[7] (minimal pair searches, context searches). The grammatical analysis comprises both qualitative methods, such as elicitation sessions with translation tasks and grammaticality judgments, and in quantitative methods, such as corpus searches and comparisons. Both programs, FLEx, where I organized my text corpus, and ELAN, provided me with Key Word In Context (KWIC) and more advanced context query possibilities. Using these options, I was able to investigate the distribution of grammatical categories and more detailed semantics. As a framework for the description of the language, I relied on an ontology of terms traditionally used in linguistic typology, which is generally referred to as Basic Linguistic Theory (BLT) (Dixon, 2009, 2011, 2012; Dryer, 2006). Whenever I deviate from this ontology and use terms differently

[5]retrieved 2020-10-12, from `https://archive.mpi.nl/tla/elan`
[6]retrieved 2021-02-12, from `https://www.fon.hum.uva.nl/praat/`
[7]retrieved 2021-02-12, from `https://software.sil.org/phonologyassistant/`

or introduce more idiosyncratic terms, I define and explain these terms to ensure that the described phenomenon can be identified correctly by the readers. Terms of syntactic units, such as “verb phrase” or “noun phrase”, are also taken from BLT. In the sketch grammar, they are used to describe language-specific units and not to make claims about universal syntactic constituencies.

2.4 Previous research

Abesabesi is an under-described language, as existing resources, especially comprehensive descriptive texts, are scarce and accessible language data has not been existent apart from wordlists and examples found within academic publications. Most publications that contain information about Abesabesi are genetic classifications (see Section 2.1.5). The data used for the majority of the classification attempts are taken from the wordlists gathered by Ibrahim-Arirabiyi (1989). He collected 400 items to compare the different Abesabesi dialects and reconstruct proto-phonemes. It remains the only lexical data source for Abesabesi dialects other than Ekiromi. Yet, as Ibrahim-Arirabiyi is not a native speaker and merely worked with lexical items, his data is faulty and contains translation errors, on which his hypotheses are based. Besides the reconstruction, his thesis contains a description of the phonology, including findings on syllable and word structure and phonological processes.

Before Ibrahim-Arirabiyi, three BA/MA theses were written about Abesabesi dialects at the University of Ilorin in Nigeria. All three theses provide a phonological overview of one Abesabesi dialect: Akpes (Akunnu) by Daramola (1984), Ekirom (Ikaram) by Raji (1986), and Oshugu (Daja) by Ayoola (1986). I did not have access to these theses.

Most of the other resources are written by Taiwo Agoyi, a native speaker from Ikaram and a lecturer at the Adekunle Ajasin University Akungba. Her master’s thesis (Agoyi, 1998) and one derived paper (Agoyi, 2001) deal with initial vowel alternations to mark plural. She noticed the differentiation between [+human] and [-human] nouns and that plural marking is dependent on this feature. While [+human] nouns change their initial vowel to /a/, [-human] nouns change it to /i/ and a small group of nouns stay unchanged. Moreover, she describes plural agreement on adjectives and demonstratives. Yet, she does not analyze these vowel alternations as noun classes but sees the similarity between the alternations in Abesabesi and those of Edoid languages and Ukaan. As languages to the west (Yoruboid, Akokoid) do not exhibit this type of vowel alternation but Edoid and Ukaan do, her conclusion is that Abesabesi and Ukaan are supposed to be subsumed in the Edoid group. In her PhD thesis (Agoyi, 2008), she describes vowel harmony in Abesabesi and provides an analysis using Optimality Theory. Her findings are four domains that trigger vowel harmony and different types of harmonies that are based on combinations of the features ATR, lowness, and roundedness. Following the use of these harmony types across the four domains, she groups the speech forms spoken in the nine settlements into four dialects (see Section 2.1.4). Two papers that derived from her thesis treat the vowel

harmony system of one dialect each: Akpes (Agoyi, 2010) and Oshugu (Agoyi, 2012b). A more detailed account of the vowel harmony system will be given in Section 3.1.4.4.

Other descriptive work includes a paper about the syllable structure of Abesabesi by Agoyi (2015a), where she explains that Abesabesi coda consonants originate in word final vowel deletion, and two papers about numerals by Agoyi (2012c) and Oyebade (2015). Both of the latter describe the composition of numerals and conclude a vigesimal system with base lexemes for 1-10, 20, and 400.

A couple of non-descriptive papers have been written by Agoyi and Oyetade. Agoyi (2013b) discusses the history of the Abesabesi-speaking settlements. Oyetade (2004) is a socio-linguistic survey on language acquisition, proficiency, domains, and attitudes of Abesabesi speakers. It attempts to assess the degree of endangerment Abesabesi is facing (see Section 2.1.6). Agoyi (2014) discusses various factors that influence the endangerment of Abesabesi. She describes language attitudes, the settlements' history, similarities between cultural traditions, and how other languages in the area threaten Abesabesi. She also presents a survey displaying the age range of existing Abesabesi speakers. As speakers under 30 years are rare, her conclusion is that Abesabesi is endangered. Finally, she lists the measurements that have been taken to maintain the languages and their impact. Agoyi (2015b) deals with language choice in public health campaigns and uses Abesabesi as a case study. The paper discusses how indigenous languages can help increase awareness about hygiene and hygienic practices to people of rural areas. Agoyi describes hygiene practices in the Abesabesi-speaking communities and provides specific vocabulary, proverbs, and a story that deal with hygiene. A locally produced history booklet (Abimboye, 2008) was published in honor of the former king of Ikaram. It describes local customs, the history of Ikaram, and mentions the language only in a paragraph while calling it a dialect of Yoruba.

Within the scopes of the Abesabesi language development committee, Agoyi published two language learning booklets. Agoyi (2012a) proposes a Yoruba-based orthography she claims to be usable by all Abesabesi dialects. Examples, however, are all taken from the Ekiromi dialect. Agoyi (2013c) is a booklet containing the Abesabesi alphabet with example words and sentences. Additionally, words and phrases of the word fields body parts and greetings are listed. It is illustrated with pictures and meant for everyone that intends to learn the language.

Prior to this project, no recorded data had been openly available. Taiwo Agoyi collected several recordings made by herself or her students. A "Talking Dictionaries"[8] project supported by the Living Tongues Institute[9] was executed by Taiwo Agoyi but has never been completed.

[8]Talking Dictionaries is a web application that comprises several dictionaries of endangered languages. The dictionary entries contain audio recordings. (retrieved 2020-06-05 from `https://talkingdictionaries.app/`)

[9]retrieved 2020-06-05 from `https://livingtongues.org/`

Chapter 3

Sketch Grammar

This chapter contains the Abesabesi Sketch Grammar. It presents a brief grammatical description of the Abesabesi language. The basis of the findings presented here is a text corpus compiled through fieldwork in Ikaram, Nigeria. Details about the composition of the corpus and about the participants that contributed to the compilation of this corpus can be found in Sections 2.3.3 and 2.3.4. The methodology behind the analysis that led to these findings was discussed in Section 2.3.6. As data collection took place entirely in Ikaram, this sketch grammar is based on data of the Ekiromi dialect only. This means that every time an aspect of Abesabesi grammar is described in this chapter, it is confined to the Ekiromi dialect. Other dialects may exhibit variation in all parts of the grammar. If dialectal variation is known, it is indicated in the text.

This sketch grammar begins with Abesabesi's phonology (Section 3.1). It then provides an overview of the sentence and its components (Section 3.2). The following sections discuss particular word classes in detail: nouns (Section 3.3), verbs (Section 3.4), and minor word classes (Section 3.5). The formation of words and derivative affixes are the topic of Section 3.6. A more detailed view on the composition of clauses is then presented in Section 3.7. Finally, Section 3.8 discusses information structure in Abesabesi.

3.1 Phonology

This section presents findings on the phonology of Abesabesi. It begins with a discussion on the phoneme inventory of Abesabesi – vowels in Section 3.1.1 and consonants in Section 3.1.2. Section 3.1.3 deals with the phonemic composition of syllables and words. The phonemic processes of vowel deletion, dissimilation, assimilation, and vowel harmony are discussed in Section 3.1.4. An overview of tone in Abesabesi is then given in Section 3.1.5. Finally, Section 3.1.6 explains the practical orthography for Abesabesi used throughout this thesis.

3.1.1 Vowels

Abesabesi distinguishes seven oral vowels. Table 3.1 contains minimal pairs distinguishing the seven oral vowels as phonemes.

/i/	/e/	/ɛ/	/a/	/ɔ/	/o/	/u/
tì ‘push’	*té* ‘exist’		*ta* ‘build’	*tɔ* ‘roast’	*to* ‘pound’	*tu* ‘be’
	je ‘see’	*jɛ̀* ‘slice’	*ja* ‘be short’		*jo* ‘dance’	*jù* ‘bury’
ímũ ‘hunger’		*ɛ́mũ* ‘money’		*ɔ́mũ* ‘bee’		

Table 3.1: Minimal pairs – oral vowels

The oral mid vowels feature an advanced tongue root (ATR) distinction ([+ATR] vs. [-ATR]), which is also essential to the vowel harmony system (see Section 3.1.4.4). Table 3.2 contains minimal pairs exemplifying this distinction. For simplicity reasons, the [+ATR] vowels are transcribed as /e/ and /o/, while the [-ATR] vowels are transcribed as /ɛ/ and /ɔ/.

/e/	/ɛ/	/o/	/ɔ/
tèn ‘follow’	*tɛ̀n* ‘join’	*ton* ‘help’	*tɔ̀n* ‘be caused by’

Table 3.2: Minimal pairs – ATR-distinction

Abesabesi also distinguishes oral from nasal vowels, which can be seen in the near minimal pairs of Table 3.3.

Vowel	Oral	Nasal
/a/	*had* ‘shine’	*hãs* ‘peel’
/ɛ/	*awɛ̀g* ‘power’	*àwɛ̃̀* ‘life’
/ɔ/	*hɔ* ‘too much’	*hɔ̃* ‘cultivate’
/i/	*ehi* ‘word’	*ehĩs* ‘empty’
/u/	*hu* ‘die’	*hũsò* ‘scare’

Table 3.3: Near minimal pairs – oral vs. nasal vowels

The five nasal vowels do not include phonemic nasalized versions of the [+ATR] vowels /e/ and /o/. Their phonetic realization, however, occurs frequently, as deleted vowels retain their nasality, which is then realized on the remaining vowel (Example 3.1).

(3.1) (a) *ɔhɔ̃ òdùg → [ɔhõdùg]*
neck leg
‘lower leg’ (ibe298-00.088)

(b) *hɔ̃ ebid → [hẽbid]*
harvest palm_fruit
‘harvest palm fruits’ (ibe350-00.042)

Summing up, Abesabesi has twelve phonemic vowels that differ in height, backness, roundness, nasalization, and tongue root position. Tables 3.4 and 3.5 show all oral and nasal vowels respectively.

	front	central	back
high	i		u
mid	e ɛ		o ɔ
low		a	

Table 3.4: Oral vowels

	front	central	back
high	ĩ		ũ
mid	ɛ̃		ɔ̃
low		ã	

Table 3.5: Nasal vowels

The high vowels /i/ and /u/ appear to have had an ATR distinction in the past (/i/ vs. /ɪ/ and /u/ vs. /ʊ/). Evidence for this claim can be taken from their behavior in respect to VH, which will be discussed in Section 3.1.4.4.

3.1.2 Consonants

Abesabesi features seven plosive consonants, which can be bilabial, alveolar, velar and labio-velar. Table 3.6 contains (near) minimal pairs that distinguish all seven plosives.

/b/	/t/	/d/	/k/	/g/	/kp/	/gb/
ba - come	ta - build	da - pour	ka - pluck	ʃi.ga - rise	kpa - block	gbà - put

Table 3.6: (Near) minimal pairs – plosives

Plosives have a voiceless-voiced distinction in all places of articulation except for the bilabial plosive. The voiceless bilabial plosive /p/ only occurs in English loan words and in one other lexical item of the corpus: *pɔm̀pɔŋ̀ɔ̀dù̀g*. Variations of this word are also used by Agoyi (2008, 2012a) as a single example for the phoneme /p/. Ibrahim-Arirabiyi (1989, p. 42) transcribes the word for 'to hear' in Ekiromi as *pas*. In my corpus, it always appears as [kpas]. This might be a transcription error because Yoruba orthography represents the phoneme /kp/ with the grapheme <p>, since the phoneme /p/ does not exist. As the phoneme /p/ exists in one lexical item, it will be analyzed as part of the phonemic inventory, but in brackets, because it is extremely rare.

Abesabesi has four nasal consonants: a bilabial, an alveolar, a palatal and a velar. Table 3.7 contains (near) minimal pairs to distinguish the four places of articulation.

/m/	/n/	/ɲ/	/ŋ/
ɛ́namà - bird	ɛna - cow	ɛɲa - red	ɛŋa - new

Table 3.7: (Near) minimal pairs – nasals

Abesabesi also features the lateral alveolar approximant /l/. It is a distinct phoneme, unlike in Yoruba, where /l/ and /n/ are allophones and distributed by a phenomenon

called nasal harmony. Here, the alveolar lateral becomes a nasal before a nasal vowel. It is difficult to prove the non-existence of a nasal harmony, as /l/ is never followed by a nasal vowel in the corpus and a certain degree of nasalization of vowels after a nasal consonant is also very common. Nevertheless, there are several lexical items where [n] precedes an oral vowel (e.g., [nàbù] 'to stand, wait', [íʃenèʃ] 'seven' or [lògìnò] 'to spoil, break')[1]. Moreover, if /n/ did not exist and [n] were only an allophone of /l/ in onsets with a following vowel, [n] could not exist in coda position. As a matter of fact, Abesabesi has plenty of words with [n] in a coda position (e.g., [ekìn] 'one', [jen] - 'to marry, own' or [àmùn] 'language'). Therefore, /l/ and /n/ are are considered to be distinct phonemes in Abesabesi.

The Yoruba locative preposition *ní*, which is subject to nasal harmony, as its vowel usually gets deleted before vowel-initial nouns, has been borrowed into Abesabesi. Unlike in Yoruba, it only appears in its oral form [lí], but Abesabesi also does not have any words that start with a nasal vowel.

The nasal harmony in other languages of the region, such as Oko (Atoyebi, 2010, p. 17), can also affect other approximants, such as /j/ and /ɣ/, which become /ɲ/ and /ŋ/ before nasalized consonants. This, again, is not the case in Abesabesi. As [ŋ] also appears in coda positions and both appear in onset position before oral vowels and [ɣ] does not even occur once as a phone in my corpus, I analyze /ɲ/, /ŋ/, /j/ and /w/ as distinct phonemes. (Near) minimal pairs can be seen in Table 3.8.

/ɲ/	/j/
ɲas - step on	jas - take
/ŋ/	/w/
ɛ.ŋa - new	à.wa - ladder

Table 3.8: (Near) minimal pairs – /ɲ/ vs. /j/ and /ŋ/ vs. /w/

This analysis is contrary to Ibrahim-Ararabiyi (1989)'s analysis, where he argues for nasal harmony in Abesabesi and considers the nasal consonants [n], [ɲ], and [ŋ] to be allophones of /l/, /j/, and /ɣ/ before nasalized vowels.

The alveolar nasal /n/ is prone to nasal assimilation (see Section 3.1.4.3). It therefore has the allophones [m], [n], [ɲ] and [ŋ] as a syllabic nasal and in coda position before other consonants. Contrary to my analysis, Ibrahim-Ararabiyi (1989) suggests that Abesabesi features nasal harmony and only lists the nasal consonant /m/ as a phoneme. He posits that [n] is an allophone of /l/, [ɲ] is an allophone of /j/ and [ŋ] is an allophone of /ɣ/. As he worked with a fairly small wordlist, his data might have lacked counterexamples to the nasal harmony.

Besides the aforementioned approximants /l/, /j/, and /w/, Abesabesi exhibits free variation of the rhotic consonant /r/, which can be realized as a tap, a trill or an approx-

[1] My transcription for nasals might not be sufficient and measuring nasalization is complicated. Recordings of examples used in this thesis can be found at ibe332-00.090, ibe121-00.109 and ibe182-00.322.

imant. This variation is also common in Ukaan (Salffner, 2010, p. 53). For simplicity reasons, I will use the grapheme <r> to transcribe the phoneme.

Abesabesi also exhibits four fricatives and one affricate. Table 3.9 displays near minimal pairs to distinguish the fricatives and the affricate.

/f/	/s/	/ʃ/	/dʒ/	/h/
efo - white	oso - rival	eʃo - tortoise	dʒo - give birth	honò - be hard

Table 3.9: (Near) minimal pairs – fricatives and affricate

There is a labio-dental, an alveolar, a post-alveolar, and a glottal fricative and a post-alveolar affricate. While the affricate is voiced, all fricatives are voiceless. The voiced affricate /d͡ʒ/ is pronounced as a post-alveolar affricate [d͡ʒ] or as a palatal plosive [ɟ] in free variation. The post-alveolar fricative is pronounced as a fricative [ʃ] or an affricate [t͡ʃ] in free variation. This latter kind of free variation is also attested in another language of the region, Oko (Atoyebi, 2010, p. 16). Agoyi (personal communication September 2019) mentions that /t͡ʃ/ and /ʃ/ are used contrastively in the older generation, but most speakers use them interchangeably. On my field trips, all efforts in contrastive elicitation failed – even with older speakers. For all lexical items containing one of the phones, both pronunciations (with [ʃ] and [t͡ʃ]) were deemed to be equally correct. Therefore, I claim that synchronically both phones are allophones of one phoneme. This loss of contrast might have been a reanalysis fostered by Yoruba phonology or that of another surrounding language. Agoyi (2008) and Ibrahim-Arirabiyi (1989), on the other hand, both list the post-alveolar fricative /ʃ/ and affricate /t͡ʃ/ as two separate phonemes.

Abesabesi also exhibits a nasalized version of the approximant /j/: /j̃/. It only occurs in the lexeme *áhoj̃* 'bone', where it probably developed from a deleted nasalized vowel. The approximant must have adopted the nasality feature and the tone. A near minimal pair is *hoj̀* 'share' vs. *áhoj̃* 'bone'. Therefore, /j̃/ is included in the phonemic inventory, but only in brackets, as it appears only once.

Lastly, a few labialized consonants are attested in Abesabesi. Labialization occurs as secondary articulation on some consonants when a succeeding rounded vowel is deleted. Section 3.1.4 describes this process in more detail. Yet, labialized consonants also occur in lexical roots, where it is difficult to trace back a deleted rounded vowel. Only some roots, such as [ɛbʷi] 'goat', expose their phonemic form in another context. [ɛbʷi] becomes [ɛbúji=na], when a 1SG-possessive clitic is added. The phonemic form is thus /ɛbuji/. For most of the roots with labialized consonants, however, this context does not reveal a deleted rounded vowel. Table 3.10 provides an exhaustive list of these roots.

Only few (near) minimal pairs of these labialized consonants and their non-labialized counterparts exist as they occur so rarely. They are listed in Table 3.11. In other words, labialization is an attested process that cannot always be traced back as the reason for labialized consonants in lexical roots. The process might be the historic cause for labialized consonants, but synchronically, they can only be treated as phonemes.

Phone	Lexical item
b^w	*ìb^{w}eȷ̀* ‘gourd’
k^w	*k^{w}àʃ* ‘clap’ (also pronounced [wàʃ]) *ìk^{w}îrì* ‘local soup’
m^w	*m^{w}à* ‘fill up’ *m^{w}a* ‘mould’ *im^{w}aŋ* ‘laughter’
f^w	*af^{w}a* ‘leaf’ (also pronounced [afa]) *if^{w}i* ‘body hair’
s^w	*ìs^{w}aȷ̀* ‘saliva’
$ʃ^w$	*iʃ^{w}ag* ‘palm wine’ *iʃ^{w}i* ‘guinea corn’
h^w	*ɛh^{w}a* ‘skin’ *h^{w}a* ‘dig’ *h^{w}àg* ‘be stuck’ *ìh^{w}àg* ‘town’s quarter’ *ɔ̀h^{w}îr* ‘rib of a palm leaf’

Table 3.10: Lexical items with labial consonants

Ibrahim-Arirabiyi (1989, p. 22) counts five phonemic labialized consonants which are not present in all dialects: /b^w/, /k^w/, /g^w/, /ʃ^w/ and /h^w/. He interprets them as single phonemic units based on his findings on Abesabesi’s syllable structure. Agoyi (2013a, p. 38) also mentions labialization of consonants but merely treats it as a process.

Summing up, Abesabesi has 29 phonemic consonants of which nine occur in only a couple of lexical items (see Table 3.12 with rare phonemes in brackets). As marginal phonemes are included in this analysis, the number of phonemes is higher than that of previous analyses. Ibrahim-Arirabiyi (1989) includes some labialized consonants as

Non-labialized		Labialized	
íbèjí	where	ìb^{w}eȷ̀	gourd
kàs	rake	k^{w}àʃ	clap
màd	who	m^{w}à	fill up
efiloŋ	feather	if^{w}i	body hair
iʃi	fly	iʃ^{w}i	guinea corn
ísaj	stone	ìs^{w}aȷ̀	saliva
had	shine	h^{w}a	dig

Table 3.11: Near minimal pairs labialized vs. non-labialized consonants

phonemes but considers certain phones as allophones, resulting in a total of 24 phonemic consonants, and Agoyi (2008, 2012b) counts 22 phonemic consonants.

	bilabial	labio-dental	alveolar	post-alveolar	palatal	velar	labio-velar	glottal
plos.	(p) b (b^w)		t d			k (k^w) g	k͡p g͡b	
nas.	m (m^w)		n		ɲ	ŋ		
fric.		f (f^w)	s (s^w)	ʃ ($ʃ^w$)				h (h^w)
affr.				d͡ʒ				
lat.			l					
app.			r		j (j̃)		w	

Table 3.12: Phonemic consonants

3.1.3 Syllables and words

Abesabesi exhibits seven possible syllable structures (see Table 3.13). The onset may be a single consonant or left empty, and the rhyme can be composed of a single or two vowels and maximally one coda consonant. Another possible syllable structure is a single syllabic nasal.

Syllabic nasals can be found in the pronoun *ǹ* '1SG' or in interjections such as *ń.kɔ̀* 'thanks' or *ǹ.bó* 'sorry'. They assimilate towards succeeding consonants and trigger an epenthetic [j] before a succeeding vowel (see Example 3.2)

Syllable Structure	Examples
N	*ǹ* '1SG', *ń.kɔ̀* 'thanks'
V	*ɛ́* 'it', *e.bo* 'dog'
VV	*àu* '3SG.LOG', *oò.sì* 'god'
CV	*ta* 'to build', *i.ɲi* 'water'
CVV	*ò.diɛ̀* 'there', *ɛ.jau* 'monkey'
CVC	*das* 'to go', *i.jel* 'year'
CVVC	*ídiàn* 'two', *máàn* 'who'

Table 3.13: Syllable structures in Abesabesi

(3.2) (a) *ǹ ke je ɛ → [ŋ̀kéjɛ́]*
1SG PRF see it
'I have seen it' (ibe331-00.210)

(b) *ǹ é kpíʃ → [ɲ̀jékpíʃ]*
1SG PROG sneeze
'I am sneezing' (ibe265-00.033)

Syllables containing a single vowel can be found in affixes and clitics such as *ɔ́* '3SG' or *è* 'NEG' and as initial syllables in nouns: *ɔ.mis* 'king'. Two consecutive vowels (or lengthened vowels) are common in pronoun affixes such as *ɔ́ɔ́* '3SG.Q', but do not occur in many noun or verb roots. Besides a couple of Yoruba loanwords, they appear in (reduplicated) compounds and in very few roots such as *ɔdau* 'harmattan' or *òono* 'locust bean'. Attested vowel combinations in native lexemes are either a lengthened vowel or one of three diphthongs: /au/, /ia/ or /iɛ/.

Phonemically, onsets are restricted to one consonant, although onset consonant clusters are frequent in rapid speech. Some examples can be seen in 3.3

(3.3) (a) *ɔ̀furɔ̀dɔ̀ → [ɔ̀frɔ̀dɔ̀]*
rooster
'rooster' (ibe005-00.060)

(b) *ani sa síkɔ́? → [anísa skɔ́]*
people 2SG.POSS how_about
'How about your family?' (ibe162-00.264)

(c) *ǹ á sà kpé oòsì bá rɔ̃̀ èsˀ lɔ́wɔ́*
1SG PROG know that God FUT help 1PL.OBJ help

→ [*ɲ̀jáskpóòsì bárèsi lɔ́wɔ́*]
'I know that God will help us.' (ibe043-00.052)

Abesabesi, along with neighboring Ukaan, stands out amongst other languages of the region, as it allows consonants in word final coda positions. Table 3.14 dispays all consonants that can appear in coda position. As it is fairly uncommon for a language of

	bilabial	labio-dental	alveolar	post-alveolar	palatal	velar
plosive			d			k g
nasal	m		n			ŋ
fricative		f	s	ʃ		
lateral			l			
approximant			r		j	

Table 3.14: Coda consonants

that region to have coda consonants, the question may arise whether these word-final consonants are in fact onsets of a deleted vowel. In order to investigate this question, these words should not be analyzed in isolation, but with other words or morphemes following them. If followed by a consonant initial morpheme, such as some possessive pronouns, they sometimes are connected with an additional vowel (see Example 3.4).

(3.4) (a) *òdùg no* → *[òdùgùno]*
leg 1SG.POSS
'my leg' (ibe001-01.145)

(b) *ètìf no* → *[ètìfino]*
hair 1SG.POSS
'my hair' (ibe073-00.130)

(c) *ɔ́ dom no* → *[ɔ́dómino]*
3SG bite 1SG.OBJ
's/he bit me' (ibe073-00.130)

This opens two possible interpretations. Either consonant-final lexemes produce an epenthetic vowel before consonants or the additional vowel is part of the phonological form and is deleted when spoken in isolation. This additional vowel is either [i] or [u], and its distribution cannot be derived from the word's form but rather depends on the lexeme. This would be indicative of a deleted vowel. However, the production of this vowel very much depends on the speaker and the speaker's rate of speech. There are many instances where the vowel does not appear. The utterances in Example 3.4 can also be pronounced without an intervening vowel, as in Example 3.5.

(3.5) (a) *òdùg no* → *[òdùgno]*
leg 1SG.POSS
'my leg' (ibe027-00.165)

(b) *ètìf no* → *[ètìfno]*
hair 1SG.POSS
'my hair' (ibe027-00.128)

(c) *ɛ́ dom no* → *[ɛ́dómno]*
it bite 1SG.OBJ
'it bit me' (ibe182-00.274)

The vowel appears to be optional and is sometimes replaced by a schwa. Especially final nasal consonants are rarely extended with a vowel. This, on the other hand, is an argument for epenthetic vowels.

Elugbe (2012, p. 3) mentions a similar case in the Edoid language Degema, which is spoken around the Niger delta. He claims that Abesabesi and Ukaan likewise developed closed syllables by dropping final consonants and posits a rule to recover the lost vowel: "A general rule is that a final /i/ or /u/ is dropped if it has the same tone as the first stem syllable or bears a low tone". Agoyi (2015a) gives a more comprehensive account on the topic by comparing data from different dialects. She presents cognates that exhibit a

Gloss	Ekiromi	Ilueni	Akpes	Oshugu
chew	*sam*	*sam*	*sami*	*sami*
send	*som*	*som*	*so*	*soli*
chase	*bur*	*bur*	*bu*	*buli*
grind	*kɔk*	*kɔk*	*kɔki*	*kɔli*

Table 3.15: Cognates with closed syllables (Agoyi, 2015a, p. 10)

closed syllable structure in the dialects Ekiromi and Ilueni and an open syllable structure in Akpes and Oshugu (an extract is listed in Table 3.15). Akpes and Oshugu cognates either have an additional vowel or drop the final consonant. She uses the data to argue that Abesabesi has a no-coda constraint and claims that closed syllables are the result of final vowel deletion. To verify the data, I looked at the cognate list provided by Ibrahim-Arirabiyi (1989) and compared all cognates where one dialect exhibits a final consonant. In fact, all varieties exhibit multiple lexical items with final consonants. Final nasals are especially common. Daja has a very high rate of deleted consonants, while Eshuku has a very high rate of additional vowels. This means a no-coda constraint is rather far-fetched.

Additionally, she presents Ekiromi and Ilueni data with final consonants where a vowel is added when followed by morphemes with an initial consonant (as in Example 3.4). This supports her claim, but she does not mention cases, where the additional vowel is not produced (as in Example 3.5).

A phonemic final vowel that was deleted is a plausible explanation for the additional vowels, especially, because the vowel quality is not triggered by stem vowels. The fact that they are optional and sometimes replaced by a schwa, however, is evocative of a reanalysis to epenthetic vowels. Even the coexistence of both processes, the deletion of word-final vowels and the production of epenthetic vowels, could be possible. Synchronically, however, these processes cannot be separated. The recovering of these vowels in elicitation was difficult, and in some cases impossible, so I did not include them in the dictionary and the phonemic transcriptions.

	without coda	with coda
CV(C)	117	138
CV.CV(C)	61	24
CV.CVV(C)	1	
CV.CV.CV(C)	17	4

Table 3.16: Word structure frequency for native, non-compounded verbs

The minimal word structure is CV. Many function words such as *gí* 'FOC' or *mí* 'COMP' and verbs such as *ʤe* 'eat' and *hɔ̀* 'cultivate' have this structure. The most frequent verb structure, however, is CVC (about 38% of all native verbs). The word structure frequency

of verbs can be seen in Table 3.16.

Nouns start with what appears to be a defunct noun class prefix (Section 3.3.2.2 containing a vowel or two. followed by the stem. The minimal noun structure is thus V.CV. The most common noun structure is V.CVC (about 37% of all native nouns). The word structure frequency of nouns can be seen in Table 3.17.

	without coda	with coda
V.CV(C)	99	116
VV.CV(C)	3	
V.CVV(C)	3	
V.CV.CV(C)	73	20

Table 3.17: Word structure frequency for native, non-compounded nouns

Like in Yoruba (Sachnine, 2014, p. 31), there is a constraint for the vowel /u/ to appear in the word initial position. This is likely to have originated in the former noun class prefixes, as all vowel-initial words are nominals. A noun class prefix consisting of the vowel /u/ might not have existed or it might have merged with another noun class prefix.

Some nouns, most of them loanwords, do not start with a vowel when spoken in isolation. Like in other surrounding languages such as Uwu (Ayere) (Allison, 2017), they are reanalyzed to have an /i/ prefix, which only occurs when preceded by other words (see 3.6).

(3.6) *ɔ́ dʒe máŋgo → [ɔ́dʒimáŋgo]*
3SG eat mango
's/he ate a mango' (ibe224-00.125)

Example 3.6 shows that the loan word *máŋgo* has the underlying structure /imáŋgo/, which can be seen when preceded by a vowel. As the first of two consecutive vowels is deleted, only the /i/ remains.

3.1.4 Processes

Abesabesi exhibits a couple of phonetic processes that alter the pronunciation of words depending on the context. The following subsections discuss vowel deletion, dissimilation, assimilation, and vowel harmony.

3.1.4.1 Vowel deletion

Vowel deletion is a very frequent process in Abesabesi, where it is used with the effect of maintaining CV structures and avoiding consecutive vowels. This means vowel deletion deals with situations when two vowels meet across morpheme boundaries: $V_1\#\#V_2$.

Abesabesi has three ways to react to this situation: no deletion, V_1 deletion, or V_2 deletion. Agoyi (2015a) claims that Abesabesi exhibits final vowel deletion in order to account for closed syllables. This was discussed in Section 3.1.3.

No deletion occurs most frequently with object/possessive clitics that consist of a single vowel: *u* '3SG.HUM', *ɛ* '3SG.NHUM', *i* '3PL.HUM' where a whole morpheme is at risk of being deleted altogether (see Example 3.7)[2]. Instead of being deleted, the first vowel is sometimes assimilated to become a lengthened vowel instead of two vowels.[3] V_1 deletion is otherwise the default and affects all possible domains such as V + N, N + N, N + Pro etc. (see Example 3.8). Exceptions are a combination of the two high vowels in the order /u/ ## /i/ and reduplication compounds (Section 3.6.4. In these two cases, V_2 deletion takes place (see Example 3.9).

(3.7) (a) *onu i* → *[onúi]*
mouth 3PL.NHUM.POSS
'their mouths' (ibe364-00.116)

(b) *òli u* → *[òlúu]*
cloth 3SG.HUM.POSS
'her/his cloth' (ibe029-00.006)

(c) *iʃo ɛ* → *[iʃéɛ]*
house 3SG.NHUM.POSS
'its house' (ibe351-00.101)

(3.8) (a) *ɔni ɛŋa* → *[ɔnɛŋa]*
person new
'stranger' (ibe001-02.217)

(b) *atɔ iʃo* → *[atíʃo]*
floor house
'house floor' (ibe029-00.156)

(c) *bò bó jù ès* → *[bòbó jès]*
3PL.HUM FUT bury 1PL.OBJ
'they will bury us' (ibe057-00.138)

(3.9) (a) *òkú iniŋ* → *[òkúniŋ]*
all thing
'everything' (ibe167-00.235)

(b) *onu iŋi* → *[onúŋi]*
mouth water
'water side' (ibe241-00.034)

(c) *ɔ́ á nu ìdʒɔ̀* → *[ɔ́ánudʒɔ̀]*
3SG.HUM PROG go farm
'S/he is going to the farm' (ibe121-00.273)

[2]A high tone between two nouns or between a noun and a pronoun, which appears in the phonetic representations of Examples 3.7-3.9, is the genitive high tone (see Section 3.1.5.2).

[3]Note that in fast speech, the lengthening disappears and only the genitive high tone (Section 3.1.5.2) is retained.

(d) *efo efo* → *[efofo]*
white white
'white' (ibe027-00.297)

(e) *ɔŋumɛ̀ ɔŋumɛ̀* → *[ɔŋúmɛ̀ŋumɛ̀]*
sweet sweet
'sweets' (ibe111-00.101)

Deleted vowels can leave a trace in the form of certain features and suprasegmentals that are retained. These processes are presented in the following paragraphs.

Labialization A deleted /u/ in V_1 position retains the [+round] feature and passes it to the preceding consonant. This results in labialized consonants (see Example 3.10). If V_2 is a rounded vowel, this process of labialization does not take place, as labialized consonants never precede rounded vowels. In this case, the /u/ in V_1 position is simply deleted. The number of labializable consonants exceeds the number of phonemic labialized consonants given in Section 3.1.2. Further phonemic labialized consonants may exist, but do not appear in the lexical database. The corpus attests labialization of all plosives (except the very rare /p/), the nasals /m/ and /n/, all fricatives ,and the approximant /j/.

(3.10) (a) *étù èsˌ* → *[étʷèsˌ]*
hat 1PL.POSS
'our hats' (ibe029-00.055)

(b) *onu ɛna* → *[onʷéna]*
mouth cow
'the cow's mouth' (ibe029-00.147)

Nasalization Deleted nasal vowels retain their [+nasal] feature and pass it on to the remaining vowel. All oral vowels can be nasalized (see Example 3.1).

Tones Vowel deletion means that a tone-bearing mora is deleted and a tone is set afloat. This process is not understood entirely, but in most cases, the tone of the deleted vowel is deleted as well. Example 3.11 shows two compounds where V_1 is deleted together with its tone, and Example 3.12 shows two instances where V_2 is deleteded together with its tone.

(3.11) (a) *LL MM* → *LMM*
èɡɛ̀ iʃo → [èɡiʃo]
door house
'the door of the house' (ibe029-00.151)

(b) *LM HM* → *LHM*
òli áje → [òláje]
fabric mother
'the mother's fabric' (ibe067-00.024)

(3.12) (a) *LH MM → LHM*
òkú iniŋ → [òkúniŋ]
all thing
'everything' (ibe167-00.235)

(b) *MML MML → MHLML*
ɔɲumɛ̀ ɔɲumɛ̀ → [ɔɲúmɛ̀ɲumɛ̀]
sweet sweet
'sweets' (ibe111-00.101)

When a low tone between two high tones is set afloat, it is not deleted but surfaces as a non-automatic downstep (see Example 3.13 and Section 3.1.5.1).

(3.13) *HL HM → H$^{\downarrow}$HM*
éwù áje → [é$^{\downarrow}$wáje]
fabric mother
'the mother's fabric' (ibe067-00.011)

3.1.4.2 Dissimilation

Dissimilation, along with VH, causes the allomorphic distribution of the pluractionalis suffix (Section 3.4.2.4). Examples are given in Table 3.37.

3.1.4.3 Assimilation

There are three types of assimilation in Abesabesi: Vowel assimilation, nasal assimilation, and vowel harmony. Vowel harmony is discussed in its own section after this one. Vowel assimilation can occur in a $V_1\#\#V_2$ environment, where no deletion takes place. This is the case when vocalic object/possessive suffixes follow a noun or verb ending in a vowel. The noun/verb's last vowel is completely assimilated to the suffix's vowel. In fast speech, only the assimilated vowel with the genitive high tone remains and the suffix vowel and tone is deleted.

(3.14) (a) *ɛ́ ka di u → [ɛ́kadúu]*
3SG.HUM.SBJ PRF tire 3SG.HUM.OBJ
'he is tired' (ibe178-02.028)

(b) *iɲi ɛ → [iɲɛ́ɛ]*
water 3SG.NHUM.POSS
'its water' (ibe351-00.116)

(c) *iniŋ edʒe i → [iniŋedʒúi]*
thing food 3PL.NHUM.POSS
'their food' (ibe351-00.108)

Vowel assimilation also occurs in compounds with the genitive high tone (Section 3.6.2). One example is *iʃáafa* 'school', which is a compound consisting of *iʃo* 'house' and *afa* 'leaf/-book'. The final vowel of *iʃo* fully assimilates with the first vowel of *afa*.

Nasal assimilation applies to /n/ in coda or nucleus positions when followed by a consonant with a different place of articulation. Only the place of articulation of /n/ is assimilated. Before bilabial, labio-dental, and labio-velar consonants, /n/ becomes [m], before palatal consonants, /n/ becomes [ɲ] and before velar consonants, /n/ becomes [ŋ]. Before alveolar, post-alveolar, and glottal consonants, /n/ is pronounced [n]. Other nasal consants are not affected by assimilation.

(3.15) (a) *bè dɛ̀n ba* → *[bèdɛ̂mba]*
3PL.S like 3PL.OBJ
'they like them' (ibe057-00.018)

(b) *ǹ je bo* → *[ǹjébo]*
1SG.S see 3PL.OBJ
'I saw them' (ibe121-00.224)

(c) *ìkpàr ìdɛ́n ka ba* → *[ìkpàrìdɛ́ŋ kaba]*
children those PRF come
'those children have come' (ibe005-00.203)

3.1.4.4 Vowel harmony

Abesabesi exhibits a rather complex vowel harmony (VH) system. As comprehensive research on this topic exists for all dialects (Agoyi, 2008, 2010, 2012b), this section will cover VH in all four dialects after presenting the findings of this research on VH in Ekiromi. Vowel harmony is a special type of assimilation where vowels are classified into two or more sets based on specific features and words usually consist of only vowels of one of those sets (root-internal VH). Vowels in affixes then assimilate to the root vowels' group in order to fit into that set (affix-VH).

Root internal VH in Ekiromi does not allow a clear distinction of separate vowel sets. However, some restrictions exist on which vowels can be in a single word. Vowel

$V_1 \backslash V_2$	i	u	e	o	ɛ	ɔ	a
i	9	4	0	1	4	4	2
u	0	0	0	0	0	0	0
e	7	3	2	5	0	0	0
o	4	10	4	6	0	0	0
ɛ	3	3	0	0	1	2	8
ɔ	1	2	0	0	2	5	1
a	1	1	2	2	3	1	4

Table 3.18: Vowel distribution in non-borrowed V.CV-words. Row heads indicate the V_1 and column heads indicate the V_2 vowels.

distribution in basic Ekiromi lexemes[4] has three peculiarities (see Table 3.18): Firstly, the [+ATR] mid vowels /e/ and /o/ never appear in a basic lexeme with [-ATR] vowels /ɛ/ and /ɔ/ (red shading). As Ekiromi features an ATR distinction only among the mid vowels, root-internal ATR-VH also seems to be restricted to the mid vowels. Secondly, the low vowel /a/ never appears directly after one of the [+ATR] mid vowels /e/ or /o/ (blue shading). It can, however, appear before them. Thirdly, no word has an initial /u/ (yellow shading). This last restriction was explained towards the end of Section 3.1.3. The combination /i/ as first and /e/ as second vowel is also not attested in basic V.CV roots (green shading). This might be a gap in the database, as there are plenty of examples of V.CV.C-roots with this combination (e.g., *ijel* 'year', or *idʒej* 'pounded yam').

Ekiromi also exhibits affix-VH. In contrast to root-internal VH, affix-VH allows a clear distinction between vowel sets (henceforth called Set I, Set II, and Set II). Two different VH-types can be distinguished, as VH in suffixes differs from VH in prefixes.

	Suffix-VH	
	Set I	Set II
i	*iɲi no* 'my water'	*ɛ́ di na* 'I am tired'
u	*ɛ́ hu no* 'it hurts me'	*iʃum na* 'my head'
e	*ɔ́ bèlè no* 's/he feeds me'	
o	*okpo no* 'my way'	
ɛ		*ɛ̀gɛ̀ na* 'my door'
ɔ		*ɔbɔ na* 'my hand'
a		*afa na* 'my leaf'
Triggered vowel	o	a

Table 3.19: Suffix-VH examples with 1SG.POSS/OBJ pronoun *no/na*

Suffix-VH is exemplified in Table 3.19 and formalized in Table 3.20. The latter displays alternations of the suffixed 1SG.POSS/OBJ-morpheme. Depending on the quality of the stem's vowels, the morpheme is realized as *na* or *no*. Like the root-internal VH, the suffix-VH shows a clear distinction between [+ATR] and [-ATR] of mid vowels. While the [+ATR] vowels /e/ and /o/ in a root trigger the suffix vowel /o/, the [-ATR] vowels /ɛ/ and /ɔ/ trigger the suffix vowel /a/. The high vowels /i/ and /u/ do not have an overt ATR distinction. Yet, they can trigger either suffix alternation, depending on the lexeme. This means, if they are the only vowels of a stem, the triggered vowel is not predictable. This is indicative of a former ATR distinction of the high vowels, which does not surface anymore. The result is high vowels with lexically specified ATR values. As it is not detectable on a phonetic level, it has to be stored on a lexical level.

[4]Basic Ekiromi lexemes in this calculation have been restricted to V.CV-words that are not borrowed. This is the only word structure where compounding can be completely precluded. The total number of items used here is 107.

	Set I	Set II
	i	i
	u	u
	e	
	o	
		ɛ
		ɔ
		(a)
Triggered vowel	o	a

Table 3.20: Suffix-VH in Ekiromi

The vowel /a/ on the other hand behaves differently. It can occur in both sets. As discussed earlier, it co-occurs in stems with the [+ATR] vowels /e/ and /o/, but only if it precedes them. These words with /a/ as V_1 and /e/ or /o/ as V_2 are in fact the only instances where /a/ is in a word of Set I. In all other instances, /a/ triggers the vowel /a/ in morphemes. This is why it is attributed to Set II, assuming that the vowel closer to the suffix is more important for suffix-VH.

The difference between the neutral vowel /a/ and the high vowels /i/ and /u/ is their behavior in roots containing only one type of these three vowels, such as *afa* 'leaf', *ta* 'build' or *iʃi* 'housefly'. While roots containing only the vowel /a/ trigger [-ATR] suffixes of Set II, roots only containing high vowels take phonetically unpredictable suffixes – their ATR value is lexically specified. For example, *afa* 'leaf' only contains the vowel /a/. Therefore, it has the phonetically predictable ATR value [-ATR] and takes the 1SG.POSS suffix *na*: *afa na*. The ATR value of the noun *iʃi* 'housefly' on the other hand is not phonetically predictable, as it only contains high vowels. Its ATR value is lexically specified as [+ATR], which is why it takes the 1SG.POSS suffix *no*: *iʃi no*. Any of these three vowels (/a/, /i/, and /u/) can appear in a root together with the mid vowels /e/, /o/, /ɛ/, and /ɔ/, where the ATR value will be determined by the mid vowels.

	Set I	Set II	Set III
	i	i	
		u	u
	e		
	o		
		ɛ	
		ɔ	
		(a)	
Triggered vowel	e	a	o

Table 3.21: Prefix-VH in Ekiromi

Suffix-VH applies to the following suffixes: the object/possessive pronouns *na/no* '1SG.POSS/OBJ', *sa/so* '2SG.POSS/OBJ', and *ba/bo* '3PL.POSS/OBJ', the gerund *i- -aǹ* and *i- -oǹ*, and the pluractionalis forms *-adiàg*, *-odiòg*, *-àd*, *-òd*, *-àg*, and *-òg*.

Prefix-VH is formalized in Table 3.21 and exemplified in Table 3.22. The latter displays alternations of the prefixed PRF-morpheme. Depending on the first vowel of the verb root, the morpheme is realized as either *ka*, *ke*, or *ko*. This type of VH behaves almost like the suffix-VH. Only /u/ with an inherent [+ATR] value does not belong to Set I

	Prefix-VH		
	Set I	Set II	Set III
i	*ɔ́ ke ɲì ègɛ̀* 's/he has opened the door'	*ɔ́ ka di iʃo* 's/he has entered'	
u		*ɔ́ ka ʃu* 's/he has run'	*ɔ́ ko hu* 's/he has died'
e	*ɔ́ ke dʒe iniŋ* 's/he has eaten'		
o	*ɔ́ ke jo* 's/he has danced'		
ɛ		*ɔ́ ka jɛ̀ ɛnam* 's/he has sliced meat'	
ɔ		*ɔ́ ka ŋɔ iŋi* 's/he has drunk water'	
a		*ɔ́ ka ba* 's/he has come'	
Triggered vowel	e	a	o

Table 3.22: Prefix-VH examples with PRF morpheme *ka/ke/ko*

[+ATR], but forms a separate set. The triggered vowels are /e/ for [+ATR] vowels (Set I), /a/ for [-ATR] vowels (Set II), and /o/ for /u/ with a lexically specified [+ATR] value.

The following prefixes adhere to prefix-VH: the 3SG.SBJ pronoun *bà/bè/bò*, the irrealis pronouns *náà/néè/nóò* '1SG.IRR', *sáà/séè/sóò* '2SG.IRR', *anáà/anéè/anóò* '1PL.IRR', *asáà/aséè/asóò* '2PL.IRR', and *báà/béè/bóò* '3PL.IRR', and almost all TAM morphemes (*á/é/ó* 'PROG', *bá/bé/bó* 'FUT', *báa/bée/bóo* 'DES', *ka/ke/ko* 'PRF', *kinà/kinè/kinò* 'PRF.NEG', *màá/mèé/mòó* 'HAB', *bá wa/bé we/bó wo* 'EPN').

Summing up, Ekiromi exhibits ATR harmony within word-roots and a constraint of /a/ after [+ATR] vowels. Moreover, suffixes adhere to an ATR harmony and prefixes to an ATR harmony with a third set consisting only of /u/.

A few CVC verbs seemingly contradict these VH types. For example, the verb *dɛ̀n* 'to like' triggers the 3PL pronoun *bè* instead of *bà* and the verb *kèd* 'to arrest' triggers *bà* instead of *bè* (Examples 3.16 a, b). Looking at vowel deletion between verbs and vowels, this can be explained. If root vowels are deleted, vowel harmony aligns to the deleted vowel's quality instead of the remaining vowel's (see Example 3.16 c). These seemingly contradictory cases are thus likely to have originated as verb-noun compounds where the underlying verb is often not retraceable.

(3.16) (a) *bè* *dɛ̀n* *na*
3PL.HUM like 1SG.OBJ
'they like me' (ibe057-00.009)

(b) *bà* *bá* *kèd* *so*
3PL.HUM FUT arrest 2SG.OBJ
'they will arrest you' (ibe281-00.198)

(c) *bè* *fe* *ɛbuj* → *[bèfɛ́bʷi]*
3PL.HUM tie goat
'they tied the goat' (ibe327-00.069)

Agoyi (2008) investigated VH in all dialects. She claims that ATR restrictions within a root are only present in the Ekiromi dialect, where the mid vowels /ɛ/ and /ɔ/ can not co-occur with their [+ATR] counterparts /e/ and /o/. The restriction of the vowel /a/, which cannot follow a [+ATR] vowel, however, applies to all dialects. She distinguishes six VH types based on the features ATR, lowness, and roundedness. Table 3.23 and 3.24 display all six types. Type 3 (ATR+low) was added in Agoyi (2012b), where she lists two different types for the Oshugu dialect. Her analysis includes four different domains: the prefixed TAM morphemes, the prefixed subject prounoun *ba/be/bɔ/bo*[5], the suffixed object/possessive pronouns *nã/nɔ̃/nõ*, *sa/sɔ/so*, and *ba/bɔ/bo*, and the nominalization circumfix *i- -anĩ/i- -enĩ/i- ɔnĩ/i- -onĩ*.

	T1: ATR(e)		T2: ATR(o)		T3: ATR+low		T4: Low	
	Set I	Set II	Set I	Set II	Set I	Set II	Set I	Set II
	i		i		i		i	
	u		u		u		u	
	e		e		e		e	
	o		o		o		o	
		ɛ		ɛ		ɛ	ɛ	
		ɔ		ɔ		ɔ	ɔ	
		a		a		a		a
Triggered vowel	e	a	o	a	ɔ	a	ɔ	a

Table 3.23: Vowel harmony types in two sets according to Agoyi (2008, 2012b)

These VH types are the base of her internal classification (see Table 3.25). Each dialect has one VH type for prefixation and one for suffixation. The only exception is the Ilueni dialect, which has the most reduced system with one single type for all remaining domains. A few points of Agoyi's analysis do not match with the findings of this thesis. Since my data is restricted to the Ekiromi dialect, additions and changes to her analysis are only verified for Ekiromi but an analogous application to the other dialects might be possible with more data. Her Types 2 (ATR(o)) and 6 (ATR+u) match the Ekiromi

[5]Agoyi's analysis is written using her transcription style. However, the tones are omitted as they are inconsistently transcribed.

	T5: ATR + round			T6: ATR + u		
	Set I	Set II	Set III	Set I	Set II	Set III
	i			i		
			u			u
	e			e		
			o	o		
		ɛ			ɛ	
		ɔ			ɔ	
		a			a	
Triggered vowel	e	a	o	e	a	o

Table 3.24: Vowel harmony types in three sets according to Agoyi (2008)

prefixation and suffixation VH-types described earlier. Yet, she attributes /i/ and /u/ only to one set in each type and does not explain the peculiar behavior of /a/.

Domain/Dialect	Ilueni	Ekiromi	Akpes	Oshugu
Prefix:TAM	ATR(e)	ATR + u	ATR + round	ATR + low
Prefix:PRO	ATR(e)	ATR + u	ATR + round	ATR + low[6]
Suffix:NMLZ	ATR(e)	ATR(o)	Low	Low
Suffix:PRO	/	ATR(o)	Low	Low

Table 3.25: Vowel harmony types in different dialects according to Agoyi (2008, 2012b)

3.1.5 Tonology

Abesabesi has three register tones: a high tone (H) (marked with an acute accent), a mid tone (M) (not marked on vowels) and a low tone (L) (marked with a grave accent).

(3.17) Three register tones

(a) High tone: *té* 'to exist'

(b) Mid tone: *ɛɲɔ* 'snake'

(c) Low tone: *òwò* 'hoe'

[6]Agoyi (2008) claims Oshugu uses the Low type in all domains although her data shows that the prefixed domains do not behave according to her Low type. Agoyi (2012b) adds the ATR + low type and says it applies to the TAM domain but does not mention the subject pronoun domain. As the data for subject pronouns in Agoyi (2008, p. 82) is also evocative of the ATR + low type, this table attributes ATR + low to the Prefix:PRO domain of Oshugu.

The low tone has three different allotones. Following an M, it is produced as a low register tone (see Figure 3.1). Following an H, it is produced as a falling tone, usually starting at a higher pitch than the preceding H (see Figure 3.2). Being the last low tone after one or more consecutive low tones at the end of an utterance, it is produced as an extra low tone (XL) (see Figure 3.3).

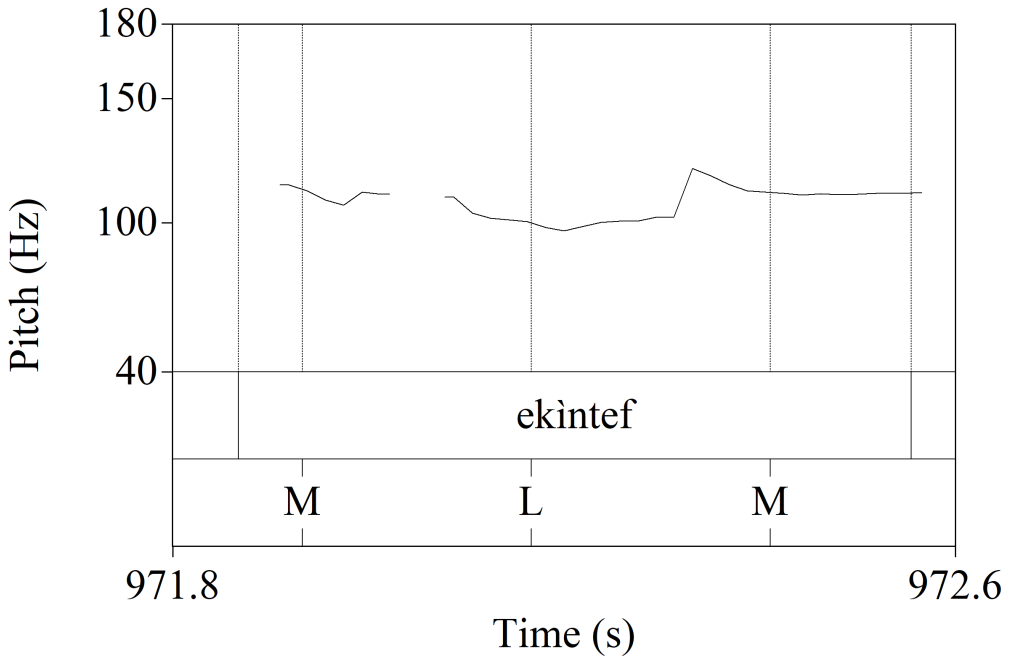


Figure 3.1: Low tone after mid tone in *ekìntef* 'eleven' (ibe121-00.113)

The tone-bearing unit (TBU) is the mora. Light syllables bear one tone, and heavy syllables with long vowels, two vowels, or coda consonants can bear two tones. However, only /j/ and nasal consonants can bear a tone in coda position. The tone borne by a coda consonant can only be L.

(3.18) Coda consonants bearing a tone

(a) *dʒom̀* 'to swear'

(b) *ahuǹ* 'nose'

(c) *ɨnadʒiŋ̀* 'shyness'

(d) *baj̀* 'to tear'

The coda consonant's restriction to low tones is probably the result of final vowel deletion and explains why Elugbe (2012) posited the rule that final /i/ or /u/ is dropped if it bears the same tone as the first stem syllable or a low tone (Section 3.1.3). Other coda consonants cannot bear a low tone of deleted final vowels. Instead, the deleted tone is set afloat to the right (see Example 3.19)

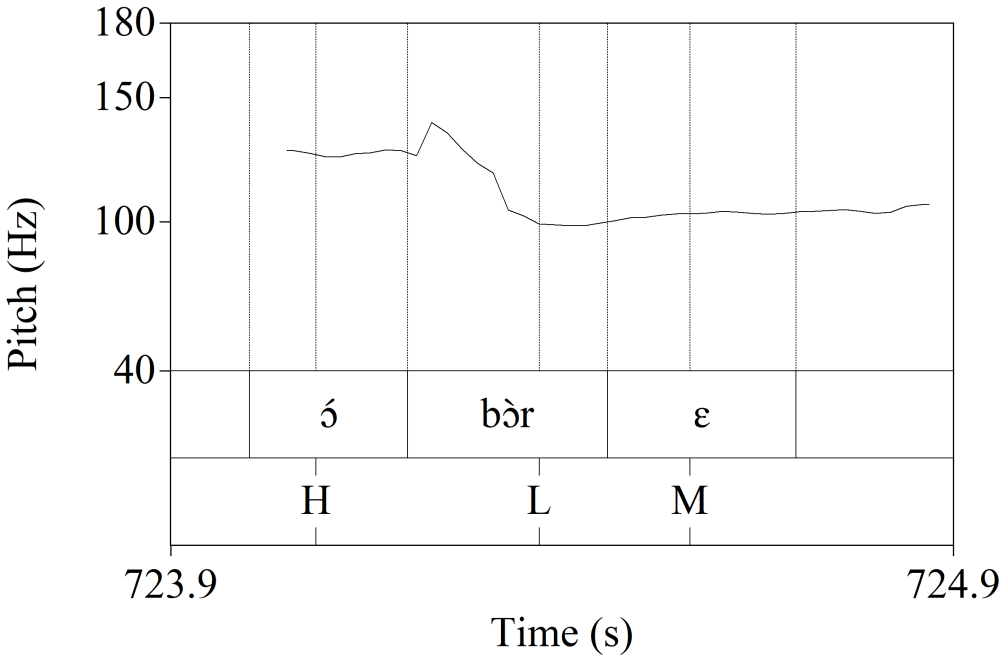


Figure 3.2: Low tone after high tone in *ɔ́ bɔ̀r ɛ* ‘s/he swallows it’ (ibe121-00.079)

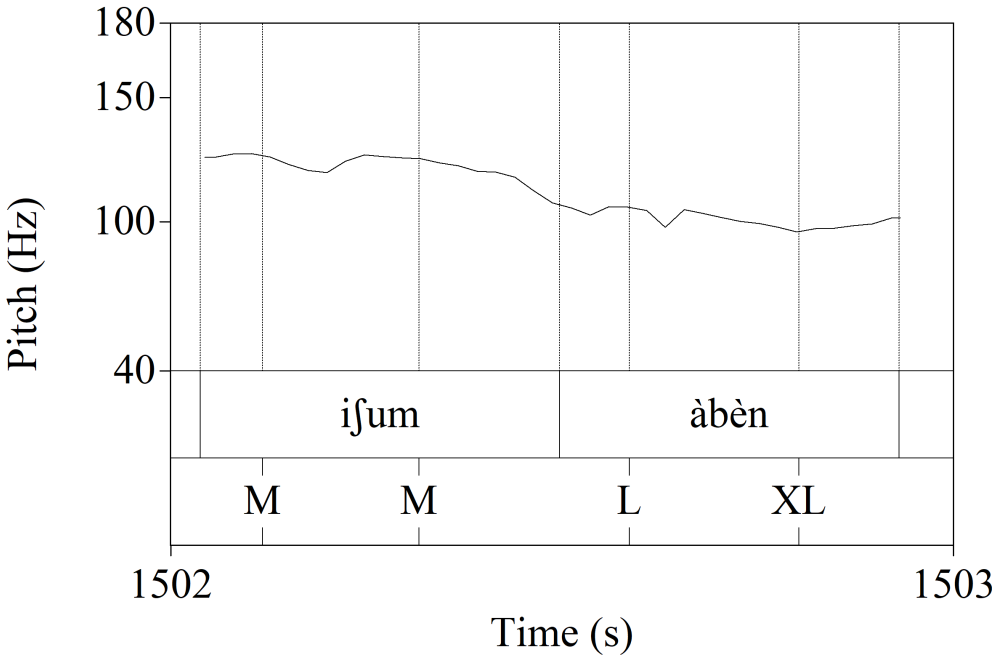


Figure 3.3: Low tone after another low tone at the end of an utterance in *iʃum àbèn* ‘your heads’ (ibe121-00.172)

(3.19) *ɔ́ wuʃ-Ⓛ ɛ deŋ → [ɔ́wuʃɛ̀deŋ]*
3SG.S dig_up 3SG.NHUM.OBJ up
's/he dug it up' (ibe331-00.574)

3.1.5.1 Automatic and non-automatic downstep

Abesabesi features automatic and non-automatic downstep. Automatic downstep (downdrift) is triggered by low tones and lowers the pitch of the following high tone. This way, the general pitch level of all succeeding tones is lowered. Example 3.20 shows an utterance where the high tone of the 3SG.HUM pronoun *ɔ́* is downstepped after a low tone. The pitch contour can be seen in Figure 3.4, where the pitch of the second high tone is substantially lower than the one of the first high tone.

(3.20) *ímèj ɔ́ ɲadum? → [ímèjˈɔ́ɲadum]*
where 3SG.HUM sit
'where did s/he sit down?' (ibe005-00.268)

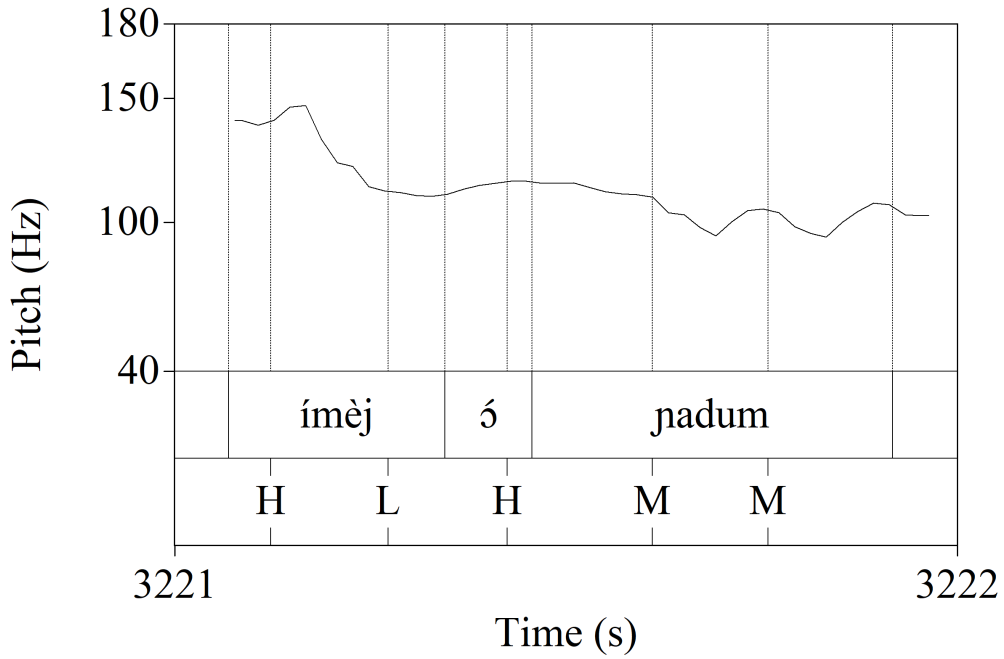


Figure 3.4: Automatic downstep in *ímèj ɔ́ ɲadum* 'where did s/he sit down' (ibe005-00.268)

This pitch lowering can occur recursively. In texts, however, the pitch often regenerates after pauses or due to prosodic pitch changes. Example 3.21 shows an utterance where the pitch is repetitively lowered on the second (*lí*) and third (ɛ̀dʒɛ́j) high tone through automatic downstep. Note that the combination of *lí* 'LOC' and *ìwùd* 'in' is often contracted to [ljúd]. Figure 3.5 displays the pitch contour.

(3.21) *mi ɔ́ gbà i lí ìwùd èdʒɛ́ĵ*
SIP 3SG.HUM put 3PL.NHUM in basket finish

→ [mɔ́gbì lj'údɛ̀dʒ'ɛ́ĵ gɛn]
'after he put them into the basket' (ibe023-00.069)

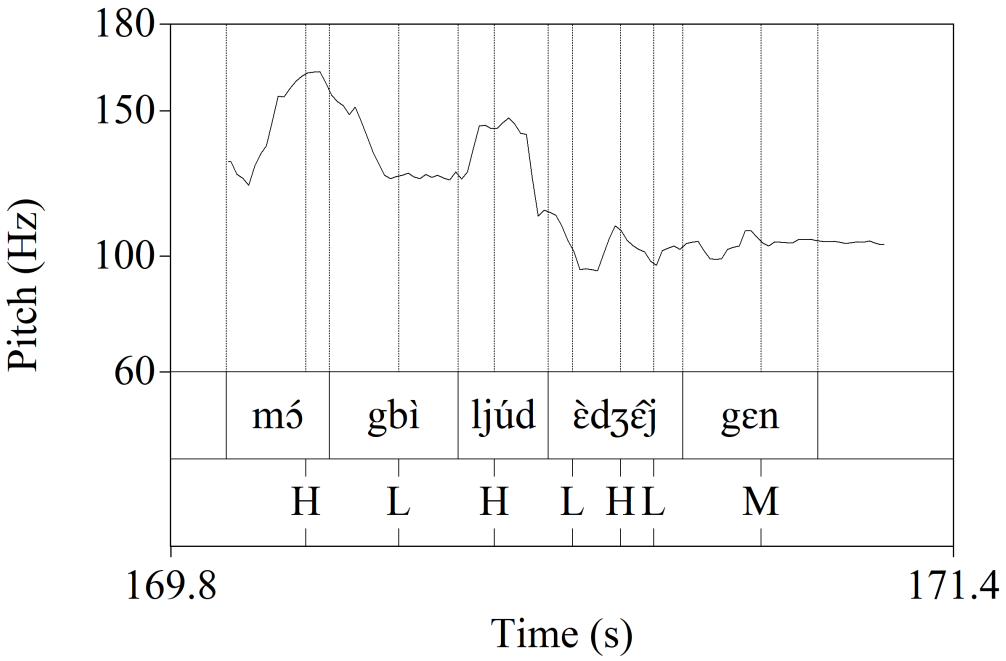


Figure 3.5: Recursive automatic downstep in *mi ɔ́ gbà i lí ìwùd èdʒɛ́ĵ gɛn* 'after he put them into the basket' (ibe023-00.069)

Non-automatic downstep occurs when low tones are set afloat and are followed by a high tone. The immediately following high tone is realized with a lower pitch as well as all following high tones. Tones may be set afloat due to deletion or grammatical high tones without segment. The non-automatic downstep does not occur in basic lexemes but can be found in complex lexemes due to vowel deletion. It is also frequent in genitive constructions and the pronoun paradigms of interrogative and conditional subject pronouns, as grammatical high tones before and after these pronouns cause the inherent low tone on some pronouns to undock and float (see Section 3.5.2.1).

(3.22) *éwù áje → [é'wáje]*
dress mother
'the mother's dress' (ibe067-00.011)

(3.23) *sáà H ba ìdʒɛ → [sá'á badʒɛ̀]*
2SG.IRR 1H come today
'if you come today' (ibe168-00.015)

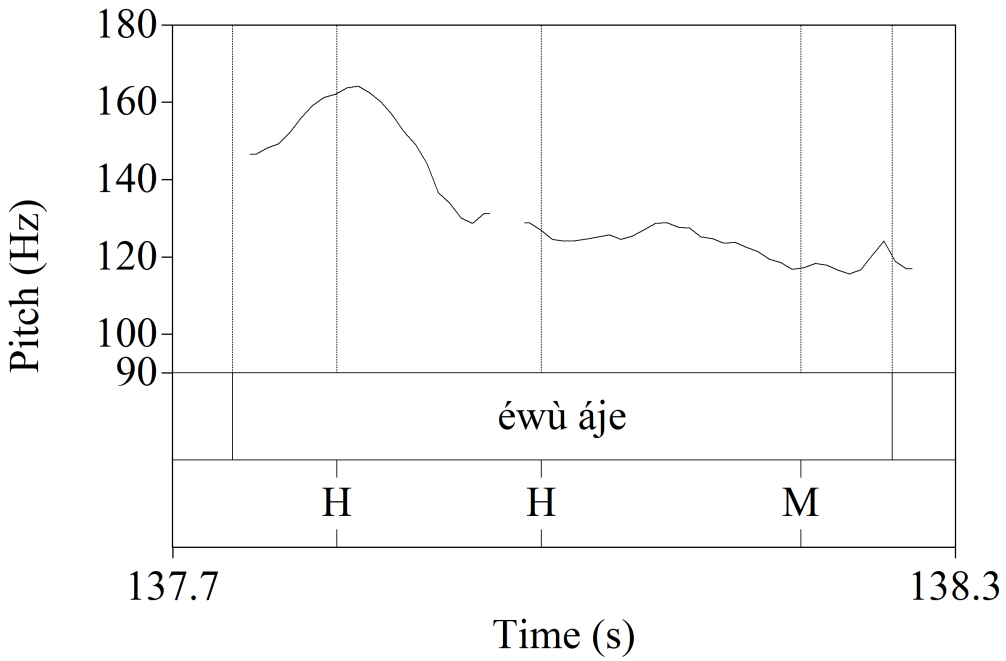


Figure 3.6: Non-automatic downstep in *éwù áje* 'the mother's dress' (ibe067-00.011)

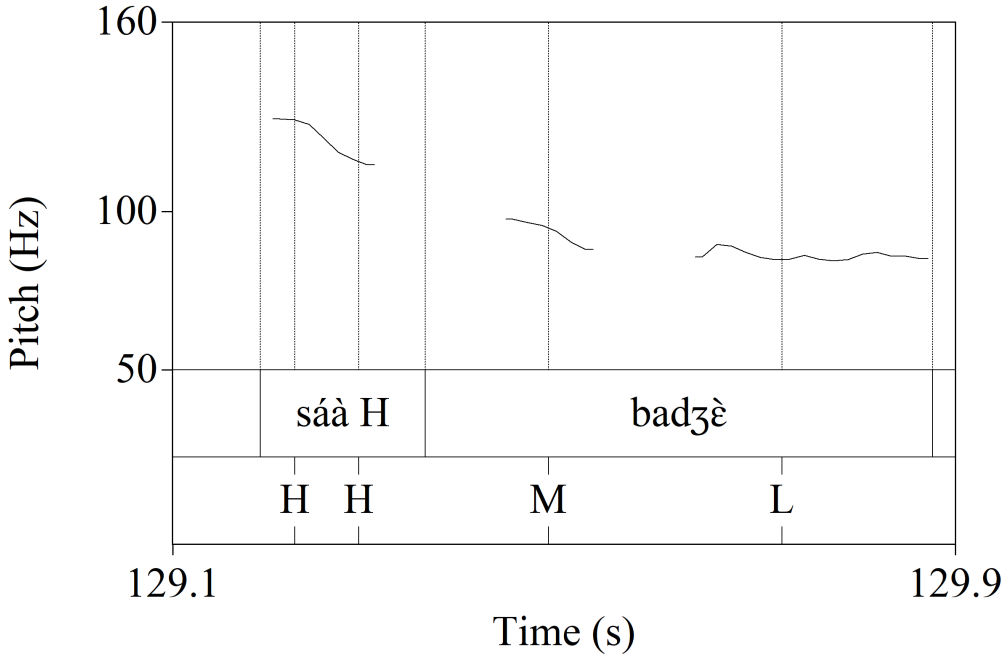


Figure 3.7: Non-automatic downstep in *sáꜜá ba ìdʒɛ* 'if you come today' (ibe168-00.015)

Example 3.22 shows a non-automatic downstep in a genitive construction. As two vowels meet, the final vowel of the first word is deleted, leaving a floating low tone between two high vowels. The second high vowel is realized with a lower pitch, as the floating low tone causes a downstep. The pitch contour can be seen in Figure 3.6. Example 3.23 shows a non-automatic downstep in a conditional clause. The conditional pronoun of the second person *sá'á* is a combination of the bound irrealis pronoun sáà and a first position high tone. As it docks to the second vowel, the low tone is set afloat and causes a downstep. The pitch can be seen in Figure 3.7.

3.1.5.2 Function of tones

The different pitch levels can be used to distinguish lexemes and to mark grammatical relations. Lexical tones have little functional load in native words. The compiled dictionary contains only 26 tonal minimal pairs (see Table 3.26). The sample size, however, is rather small, as only 824 items in the dictionary are not borrowed.

H		M		L	
íniŋ	four	iniŋ	thing		
		fu	cook	fù	wash
		ʃerè	to run around	ʃèrè	to produce
		otu	thief	òtù	farm plot
		idʒɛ̀	bone	ìdʒɛ	today
		adɔm	storage	àdɔ̀m	moon
sá	to end	sa	2SG.OBJ/POSS	sà	to know
		ɲum	to wring	ɲum̀	to be sweet
		ton	to meet	toǹ	to carry
		hoj	grow	hoj̀	to share

Table 3.26: Tonal minimal pairs

Tones on basic native lexical items are mostly M and L. If borrowed, complex, and functional items are excluded from the discussion[7], H is found on only 7% of lexical items. This could be indicative of a former two-tone system similar to the one of Ukaan (Salffner, 2009, p. 91) and a few Edoid languages of the region (Elugbe, 1989, pp. 97-99). High tones are found on most function words and some loan words. Otherwise it is reserved for grammatical functions. The high tone is a poly-functional toneme marking different grammatical relations: The first position high tone (1H), the genitive high tone (GenH), the relative high tone (RelH), the locative high tone (LocH), and the interrogative high tone (QH). Their form and function are discussed in the following paragraphs. Note that the remarks on occurrence and functions are to be treated with care, as no in-depth

[7]Function words are excluded because the high tone that most of them carry might be a grammatical tone and not part of the lexeme.

empirical research could be undertaken. They should not be seen as generalizations but as observations and impulses for future research.

First position high tone (1H) The 1H occurs at the end of the first phrase of a sentence. This is usually the subject of the first clause in a sentence, no matter if the clause is embedded or not. However, not only subjects but also fronted interrogative pronouns can be marked with the 1H. The high tone affects the last mora of the first phrase. It merges with a high tone (Example 3.24 a), replaces a mid tone (Example 3.24 b), and sets a low tone afloat, causing a downstep (Example 3.24 c). If the affected noun ends in a consonant, an epenthetic /i/ is added to bear the high tone (Example 3.24 d).

(3.24) (a) *ìwé H ga → [ìwé ga]*
book 1H be_big
'the book is big' (ibe363-00.022)

(b) *ebo H miʃ → [ebó miʃ]*
dog 1H sleep
'the dog sleeps' (ibe363-00.014)

(c) *étù H ga → [étˡú ga]*
hat 1H be_big
'the hat is big' (ibe363-00.011)

(d) *ìkpàr H ka das ìlúdɔ̀tũ → [ìkpàrí kadas ìlúdɔ̀tũ]*
children 1H PRF go Iludotun
'the children have gone to Iludotun' (ibe363-00.009)

The 1H does not occur in sentences with defunct verbs (3.25 a), such as *gó* 'to be', *té* 'to exist', or *síkɔ́* 'how about?' (Section 3.4.4.3). It also does not occur in negated sentences (3.25 b), sentences with imperatives (3.25 c), or focus-fronting (3.25 d).

(3.25) (a) *ɔhunɛ gó → [ɔhunɛ gó]*
tree be
'it is a tree' (ibe005-00.296)

(b) *ebo miʃ è → [ebo miʃìjè]*
dog sleep NEG
'the dog does not sleep' (ibe363-00.043)

(c) *àn ʃì mí à das → [ànʃì máàdas]*
2PL let COMP 1PL go
'let us go' (ibe017-00.271)

(d) *ònì gí bɔ̀r atɔ → [ònì gí bɔ̂ratɔ]*
1SG.IDP FOC hoe ground
'it is me who hoed the ground' (ibe128-00.005)

Negative focus, however, exhibits the 1H on the focused phrase although it does not occupy the first position (Section 3.8.2).

(3.26) *éɛ̀ tu àbès H bɔ̀r atɔ è → [éɛ̀ tàbèsí bɔ̀ratɔè]*
it.NEG be 1PL.IDP 1H hoe ground NEG
'it is not us who hoed the ground' (ibe128-00.038)

Bound subject pronouns also take the 1H. Apart from the 3SG pronouns, which bear a high tone, all bound subject pronouns bear a low tone. As low tones cannot be set afloat at the beginning of a word, the high tone is pushed to the following element (see Example 3.27). Since the subject pronouns are bound, the following element is either a TAM marker or a verb.

(3.27) *ɔ̀ H ka ba → [ɔ̀kába]*
2SG 1H PRF come
'you have come' (ibe121-00.183)

Genitive high tone (GenH) The GenH is used on modified nouns or verbs. It occurs on nouns that are modified by other nouns or possessive pronouns (3.28) and on verbs before object pronouns (3.29). Verbs with incorporated objects are also marked with the GenH when taking an additional object (incorporation is discussed in Section 3.6.3). Example 3.29 b displays an utterance where the verb *jotu* 'steal' (the verb *ju* 'act as' has incorporated the noun *otu* 'thief') takes the additional object *ɛmũ* 'money'.

(3.28) Genitive high tone on nouns

(a) *iʃo H ɛna → [iʃéna]*
house GenH cow
'the cow's house' (ibe029-00.146)

(b) *ɛmũ H ba → [ɛmű́ba]*
money GenH 3PL.POSS NEG
'their money' (ibe363-00.043)

(3.29) Genitive high tone on verbs

(a) *ijes H ó hu H no → [ijesí jóhúno]*
body 1H PROG hurt GenH 1SG.OBJ
'I am in pain' (ibe331-00.355)

(b) *ɔ́ H ko jotu H ɛmũ èdén → [ɔ́kojotém^{w}ɛdèń]*
3SG 1H PRF steal GenH money that
'he has stolen that money' (ibe331-00.430)

GenH occurs at the end of the modified word. It only occurs if the first word ends on a mid tone and the second begins with a mid tone (see Examples 3.28 and 3.29 above). Example 3.30 shows instances where either the first word does not end on a mid tone (3.30 a) or the second word does not start with a mid tone (3.30 b). This suggests that the GenH is not a toneme, but a process of mid tone raising.

(3.30) (a) *bò H bó jù no* → *[bòbójùno]*
3PL.HUM 1H FUT bury 1SG.OBJ
'they will bury me' (ibe057-00.135)

(b) *iʃum ès* → *[iʃumès]*
head 1PL.POSS
'our heads' (ibe005-00.177)

Other than the 1H, there is no epenthetic /i/, but the high tone is placed on the last vowel. If the first word ends in a tone-bearing consonant, it keeps the tone, while a mid tone on the preceding vowel is raised to a high tone (Example 3.31 a). As described in Section 3.1.3, an additional vowel may appear when coda consonants are followed by another consonant (Example 3.31 b). These vowels are never raised by the genitive high tone.

(3.31) (a) *iʃum H ba* → *[iʃúm̄ba]*
head GenH 3PL.POSS
'their head' (ibe121-00.179)

(b) *ɛʃag H na* → *[ɛtʃágina]*
crab GenH 1SG.POSS
'my crab' (ibe001-01.104)

Relative high tone (RelH) The RelH is similar to the GenH in that it is also carried by a modified noun, but RelH exhibits a slightly different behavior. The high tone docks onto a noun that is modified by a relative clause. On vowel-final nouns, it replaces the tone on the last vowel. In contrast to the GenH, this also affects low tones (see Example 3.32).

(3.32) (a) *àkɛtɛ̀ H mi ɔ́ fe*
woven_hat RelH REL 3SG.HUM wear

→ [àkɛtɛ́ mɔ́ fe]
'the woven hat he wears' (ibe041-00.062)

(b) *ǹ H dɛ̀n ɔni H mi sɛmɛ̀ na*
1SG 1H like person RelH REL greet 1SG.OBJ

→ [ǹdɛ̂n onínsɛmɛ̀na]
'I like the person that greeted me' (ibe208-00.002)

If the noun has a final consonant, an epenthetic /i/, which bears the RelH, is attached to the right end of the noun (see Example 3.33). Relative clauses are discussed in Section 3.7.4.2.

(3.33) *owos H mi das ìbùrɛ H ɛ*
man RelH REL go time GenH 3SG.NHUM.POSS

→ [owosínd̄as ìbùrɛ́]
'the man that left the other time' (ibe229-00.112)

Locative high tone (LocH) Phrases denoting a location are usually situated at the end of a clause. They consist of the LocH and a noun or pronoun that indicates a location. The tone precedes the noun and therefore docks onto the initial vowel. If the noun or pronoun starts with a consonant, an epenthetic /i/ appears where the H can doc (Example 3.34 c)

(3.34) (a) *owos ekìn tu H òdiɛ̀ → [oosekìń tódiɛ̀]*
man one be LocH there
'one man is there' (ibe041-00.002)

(b) *àdúgbò ɔmis gí bò bò dʒo H no H èkiròm*
quarter king FOC they PST give_birth GenH 1SG.OBJ LocH Ikaram

→ [àdúgbɔ̀mis gí bòbòdʒóno ékiròm]
'the quarter of the king was where I was born in Ikaram'
(ibe119-00.003)

(c) *H mèj H ɔ̀ tu? → [ímèjɔ́ɔ̀tu]*
LocH where 1H 2SG be
'where are you?' (ibe017-00.329)

Marking locative phrases with an initial high tone is common in the area. Atoyebi (2010, p. 57) mentions a high tone in Oko and traces it back to a deleted high tone vowel. The LocH in Abesabesi is often replaced by the Yoruba preposition *ní*, which is borrowed in its non-nasal form *lí*, or by the preposition ʃí (Section 3.2.2.3).

Interrogative high tone (QH) Polar questions resemble affirmative clauses but are introduced by a high tone, the QH. The tone docks onto the first word's first mora. Example 3.35 displays three polar questions introduced with the QH. The subject of Example 3.35 a starts with an L, the subject of Example 3.35 b with an M, and the subject of Example 3.35 c with an H.

(3.35) (a) *H òdè H ga? → [ódˡé ga]*
QH stool 1H be_big
'is the stool big?' (ibe363-00.055)

(b) *H ebo H miʃ? → [ébó miʃ]*
QH dog 1H sleep
'does the dog sleep?' (ibe363-00.053)

(c) *H étù H ga? → [étˡú ga]*
QH hat 1H be_big
'is the hat big?' (ibe363-00.052)

In this last example, the subjects start with an H, which merges with the QH. The polar question would thus not be distinguishable from its declarative equivalent (Example 3.36). However, the pitch in polar questions starts much higher, which might be

a prosodic phenomenon. This can be seen in the pitch contours of the utterances from Example 3.35 c and 3.36 (see Figures 3.8 and 3.9).

(3.36) *étù H ga → [ét$^{\downarrow}$ú ga]*
hat 1H be_big
'the hat is big' (ibe363-00.011)

As bound subject pronouns in polar questions receive the QH and the 1H from both sides, an inherent low tone is set afloat resulting in a downstep. Additionally, pronouns are lengthened. In this work, all of these pronouns are transcribed together with both high tones and the downstep to separate them from indicative pronouns. Moreover, they are treated as a separate paradigm (see interrogative pronouns in Section 3.5.2.1). Example 3.37 shows the composition of interrogative subject pronouns for the second and third person and how they are transcribed.

(3.37) (a) *H ɔ̀ H → ɔ́$^{\downarrow}$ɔ́ '2SG.Q'*
QH 2SG.SBJ 1H

(b) *H ɔ́ H → ɔ́ɔ́ '3SG.Q'*
QH 3SG.SBJ 1H

3.1.6 Orthography

The practical orthography used in this research is phonemic with a few rules to ensure consistency and simplicity. IPA symbols are used as graphemes. Diacritics are only used when they mark a phonemic feature. For instance, labialization is only written where it appears in root morphemes, ties to mark double articulation are omitted, but tone marks and the nasalization tilde are used. As the orthography is phonemic, only one grapheme is used for allophones: <r> for all rhotic sounds and <ʃ> for the allophones [ʃ] and [t͡ʃ]. Vowel harmony is represented as it is pronounced instead of writing archiphones. Assimilation is not represented in the orthography and the original phonemes are transcribed. Vowel lengthening is represented by two vowel graphemes. The more drastic rules concern wordhood and tones. These rules have been adapted to facilitate the transcription of large amounts of data. Phenomena that are thus not transcribed, such as the five grammatical tones, have been described in the previous paragraphs so that the information is not lost to future researchers. In the sense of language documentation, a more detailed description could be added in later stages if research extends to these phenomena.

Wordhood Although phonemically the language is rather agglutinative, I am writing affixes and clitics as separate words. This decision is inspired by orthographies of other Nigerian Benue-Congo languages, especially the one for Yoruba. The biggest advantage is that complicated sandhi processes, such as vowel deletion and its effects on tone and

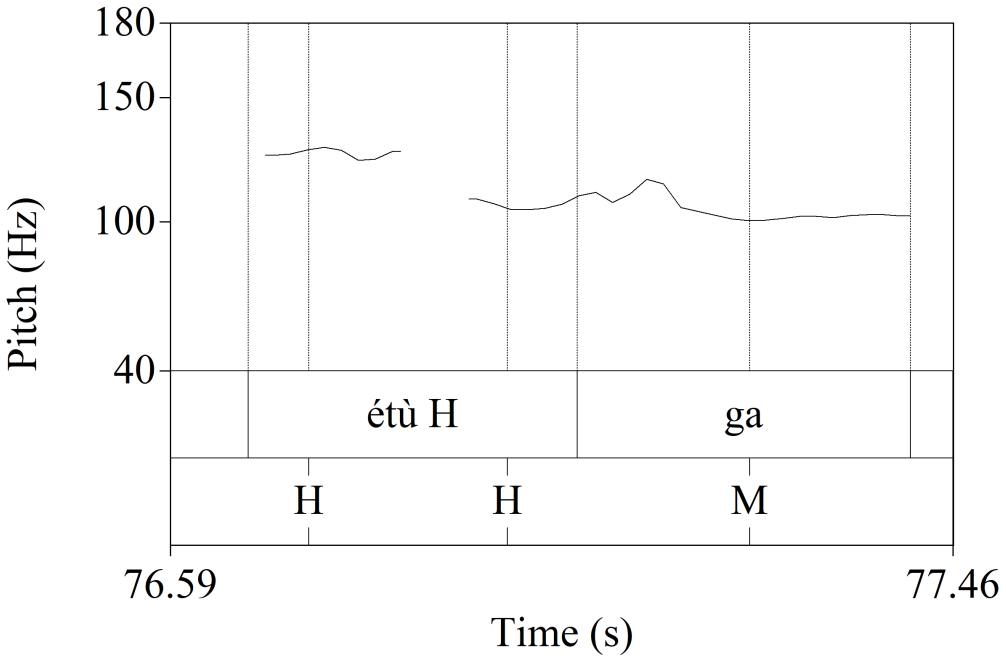


Figure 3.8: Affirmative *étù H ga* [étˈú ga] ‘the hat is big’ (ibe363-00.011)

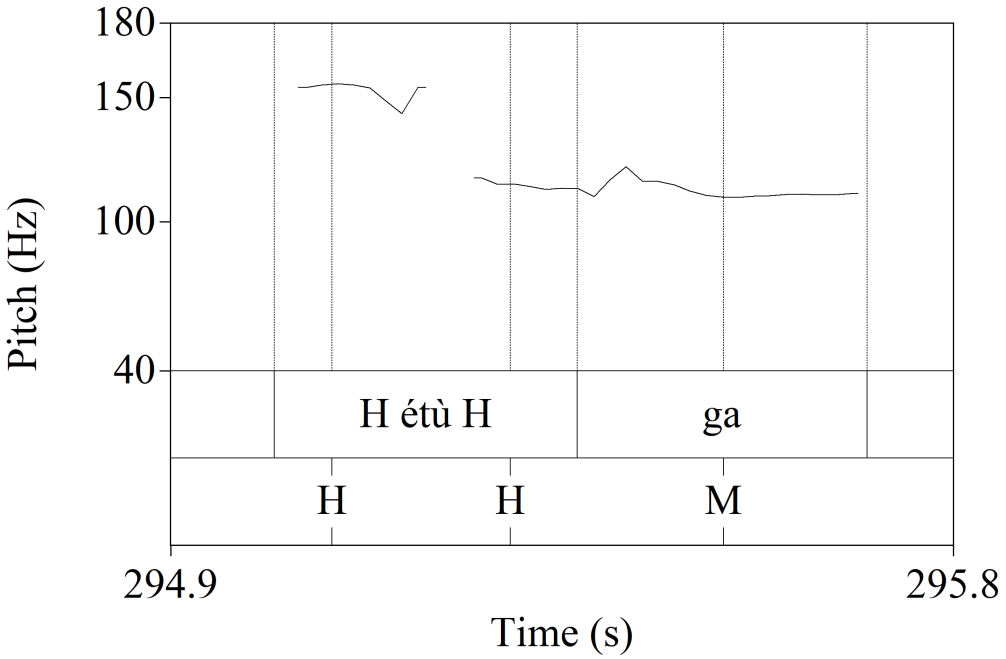


Figure 3.9: Interrogative *H étù H ga* [étˈú ga] ‘is the hat big?’ (ibe363-00.052)

other suprasegmentals, do not have to be transcribed. Although not transcribed as affixes, all categories described in the sections 'Nominal morphology' (Section 3.3.2) and 'Verbal morphology' (Section 3.4.2) are in fact analyzed as affixes.

Tones The three register tones are marked with the acute accent (high tone) and the grave accent (low tone). The mid tone is unmarked. Nasals and the approximant /j/ in coda position can bear a low tone, which is why they can also have an accent mark. If the coda consonant is unmarked, it means that it bears the same tone as the syllable's nucleus. Syllabic nasals can bear all three tones and are marked in the same way as the vowels. Since the grammatical tones described in the previous section and floating tones can cause shifts in the tones of lexical roots and other processes, which cause a labor-intensive and confusing transcription, they are not transcribed. Some examples in the following sections, however, will transcribe grammatical high tones as H when the tone is essential for a construction. As non-automatic downsteps only occur due to deleted vowels or grammatical tones, they do not have to be transcribed. The only exception are the paradigms for interrogative and conditional subject pronouns. These are both transcribed with the grammatical tones and a downwards arrow for a non-automatic downstep to differentiate them from the indicative paradigm. Automatic downsteps are not marked at all.

3.2 The sentence - an overview

This section gives an overview of the sentence in Abesabesi, its structure and its components. Sentences can consist of one or multiple clauses. A sentence consisting of only one clause is called a basic sentence and a sentence containing two or multiple clauses is called a complex sentence. While this section is merely an overview to familiarize the reader with the basic structure and the components of sentences, everything presented here will be discussed in detail in the respective sections within the sketch grammar.

3.2.1 The clause

Abesabesi clauses minimally contain a subject and a predicate.

$$SBJ + PRED$$

Except for imperatives, subjects must be marked overtly. They are expressed as NPs or as bound subject pronouns, while predicates can be verbs or defunct verbs. Defunct verbs do not have the full range of verbal features. For example, they do not trigger the 1H and they cannot occur with all TAM markers (Section 3.4.4.3). Example 3.38 displays a minimal clause with a defunct verb (a) and one with a prototypical verb (b).

(3.38) (a) *SBJ* *defunct V*
òní gó
1SG.IDP FOC
'it is me' (ibe346-00.140)

(b) *SBJ* *V*
ebo miʃ
dog sleep
'the dog slept' (ibe363-00.014)

Mono-verb clauses can be distinguished from multi-verb clauses. While mono-verb clauses only contain one verb, multi-verb clauses, such as verb serializations or clauses with auxiliary verbs, contain two or more verbs (see Section 3.4.5.2). A mono-verb clause can take up to three arguments, depending on the verb's valency. Examples in 3.39 display mono-verb clauses with one, two, and three arguments.

(3.39) (a) *SBJ* *V*
atɔ hud
ground be_dry
'the ground is dry' (ibe224-00.004)

(b) *SBJ* *V* *PO*
ìmúsù dʒe ehu
cat eat rat
'cats eat rats' (ibe178-00.153)

(c) *SBJ* *V* *PO* *SO*
ǹ bá ki sa ɛmũ
1SG FUT give 2SG.OBJ money
'I will give you money' (ibe190-00.471)

Arguments are marked through word order. As Abesabesi has an SVO word order, the subject occupies the preverbal position and the objects occupy postverbal positions. The arguments thus have the order subject > primary object > secondary object. However, as multi-verb clauses are more common than ditransitive verbs, the word order of the objects depends on the order of the verbs (see Section 3.4.4.2). A detailed discussion of Abesabesi's argument structure is given in Section 3.7.1.

The maximal clause consists of a subject, multiple verb phrases, multiple adjuncts, and multiple clause-final particles.

$$SBJ + VP* + ADJUNCT* + PART*$$

As the valency of a predicate can be extended through additional verbs, a multi-verb clause could theoretically have more than three arguments. This corpus, however, does not contain any clause with more than three arguments. Clauses with three arguments are mostly multi-verb clauses, as only a few ditransitive verbs exist. Example 3.40 displays a multi-verb clause with three arguments.

(3.40)

SBJ	*V*	*PO*	*V*	*PO*	
ǹ	ton	u	kèd	ɔban	u
1SG	meet	3SG.OBJ	carry	child	3SG.POSS

'I helped her to carry her child' (ibe073-00.173)

There are a few clause types which do not obey this default word order or have fronted phrases. These are embedded clauses, such as complement or relative clauses, focus constructions, and left dislocated clauses.

3.2.2 Phrase types

We distinguish three types of phrases in Abesabesi: noun phrases (NPs), verb phrases (VPs), and prepositional phrases (PPs). All phrases are syntactic categories that are defined by the word class of their lexical head that may or may not occur with certain dependent words.

3.2.2.1 Noun phrase

NPs have nouns or pronouns as their lexical head. They may be modified by other nouns, property nouns, demonstratives, numerals, possessive pronouns, or relative clauses. NPs function as arguments or adjuncts and can be dependents of other NPs or PPs. Example 3.41 shows different kinds of NPs: a minimal NP (3.41 a), an NP with a noun as modifier (3.41 b), a demonstrative pronoun as a head (3.41 c), and a heavy NP (3.41 d).

(3.41) (a)

N
ìmúsù
cat

'a cat' (ibe292-00.444)

(b)

N	*N*
ògbegel	fűmi
car	<NAME>

'Funmi's car' (ibe073-00.020)

(c)

DEM	*PN*
èéni	ɛdug
this	big

'the big one' (ibe292-00.376)

(d)

N	*PRO*	*NUM*	*PN*	*[*		*]$_{REL}$*	*DEM*
iʃo	no	ísas	iŋa	mi	ǹ	ta	ìdín
house	1SG.POSS	three	new	REL	1SG.SBJ	build	these

'my three new houses that I built' (ibe292-00.420)

All modifiers follow the head. Possible modifiers within an NP and their word order are discussed in Section 3.3.4.

3.2.2.2 Verb phrase

A VP must have a verb as its head. It optionally includes TAM prefixes and object arguments. As Abesabesi features a couple of auxiliary verbs that are either used to modify an event or to add participants, a VP may contain multiple verbs.

(3.42) (a)

SBJ	*[TAM*	*V*	*OBJ]$_{VP}$*
ebo	ke	dʒe	iniŋ
dog	PRF	eat	thing

'the dog has eaten' (ibe363-00.070)

(b)

SBJ	*[TAM*	*PreV*	*OBJ*	*V*	*OBJ]$_{VP}$*
ǹ	ka	ton	u	ʃa	ɛ̀nì
1SG	PRF	meet	3SG.OBJ	say	matter

'I have talked to her/him' (ibe203-00.319)

(c)

SBJ	*[TAM*	*PreV*	*V]$_{VP}$*
ɔ́	ka	jànà	ʃiga
3SG	PRF	do_early	rise

's/he has gotten up early' (ibe331-00.396)

3.2.2.3 Prepositional phrase

Only two prepositions exist in Abesabesi: *lí* - marking a location and *ʃí* - marking a location or goal. While consultants stated that *lí* is loaned from Yoruba (ní)[8], they confirmed that the Abesabesi equivalent is the locative high tone (LocH). *lí* can be substituted by *ʃí* when used to mark a location. *ʃí* can also be used to mark a goal. In this function, *ʃí* closely resembles the Yoruba preposition sí. In fact, Yoruba constructions containing *sí* are usually borrowed replacing *sí* with *ʃí*: Yoruba *yàtọ̀ sí* 'apart from/be different from' becomes Abesabesi *jàtɔ̀ ʃí*. Postpositions or circumpositions do not exist.

(3.43) (a)

PREP	*N*
lí/ʃí	atɔ
at	ground

'on the ground' (ibe167-00.048, ibe292-00.601, ibe292-00.414)

(b)

PREP	*N*
ʃí	atɔ
to	ground

'onto the ground' (ibe229-00.121)

3.2.3 Word Classes

Abesabesi offers a formal distinction between two major classes: vowel-initial and consonant initial words. This separation is probably a relic of a former noun-class system with vowel prefixes (see Section 3.3.2.2) which resulted in a separation of nouns and

[8]Consult Section 3.1.2 for a discussion on nasal harmony and why ní is borrowed as lí into Abesabesi.

noun-like words from verbs. Henceforth, I will call the vowel-initial words consisting of nouns and noun-like words "nominals" and the consonant-initial words "verbs"[9]. The major syntactic distinction between these two broad classes is that nominals can head NPs and verbs can head VPs.

Nominals consist of prototypical nouns and their subclasses, demonstratives, and pronouns. Besides their initial vowel and their ability to head an NP, which distinguishes them from verbs, nominals also feature a human/non-human distinction, which is a left-over from the former noun-class system. This distinction might be inherently determined (such as in prototypical nouns) or marked by agreement (such as on property nouns or pronouns). The word class of nouns is discussed in Section 3.3 along with nominal morphology, non-prototypical nouns, and the noun phrase. Determiners are discussed in Section 3.5.1 and Pronouns in Section 3.5.2. Verbs are consonant-initial, and can head VPs. They are discussed in section 3.4 along with verbal morphology, TAM categories, non-prototypical verbs, and verb phrases. The word class of adverbs might have developed from nouns as they have an initial vowel. They cannot, however, be categorized as nominals, because they cannot function as core arguments and do not head NPs. A detailed discussion can be found in Section 3.5.3. Lastly, the particle category encompasses all little words that are purely functional and cannot be assigned to one of the other categories. They are discussed in Section 3.5.4

- Nominals
  - Nouns
    - Prototypical nouns

      e.g., *ìtɔ̀j* 'grass', *ɔ̀jɛ̀* 'music', *èkiròm* 'Ikaram'
    - Property nouns

      e.g., *ɛŋa* 'new', *ɛmiŋɛ* 'small', *efo* 'white'
    - Numerals

      e.g., *ekìn* 'one', *íʃánas* 'six', *ànaàniŋ* 'eight'
    - Function nouns

      e.g., *òʃum* 'back/behind/after', *òkú* 'all'
    - Deverbal nouns

      e.g., *ikòoǹ* 'falling', *idàsaǹ* 'going', *ohu* 'death'
  - Demonstratives

    e.g., *èéni* 'this', *èdén* 'that', *ìdín* 'these'
  - Pronouns

    e.g., *àu* '3SG.IDP/LOG', *èŋeŋès* 'ours', *ɔńtɔn* 'anybody'

[9]In fact, there are a few nouns with initial consonant, mostly loan-words. However, as they are reanalyzed as words with deleted initial /i/, they are seen as words with an underlying initial /i/. For a short discussion, see Section 3.1.3

- Verbs
  - Prototypical verbs

    e.g., *miʃ* 'sleep', *jotu* 'steal', *tì* 'push'
  - Property verbs

    e.g., *honò* 'be hard', *ɲa* 'be red', *ʃen* 'be good'
  - Defunct verbs

    e.g., *gó* 'it is', *té* 'there is', *sík͡pɔ́* 'how about?'
  - Auxiliary verbs

    e.g., *jànà* 'early', *fàŋ* 'CPOR/SUP'
- Adverbs

  e.g., *èhen* 'now', *òden* 'there', *ìd͡ʒɛ* 'today'
- Particles

  e.g., *gí* 'FOC', *è* 'NEG', *mi* 'REL'

Property nouns, demonstratives, pronouns, defunct verbs, and particles are closed word classes. All the other classes are open as they contain loan-words. Deverbal nouns are the only derived word class.

3.3 Nouns and the noun phrase

This section investigates the noun and its features (Section 3.3.1), the morphology of nouns (Section 3.3.2), its non-prototypical sub-classes (Section 3.3.3), and the noun phrase 3.3.4.

3.3.1 The noun

Nouns refer to a broad range of entities or experiences in the world. As they cannot be classified in terms of their meanings, a prototype will be morphologically and syntactically defined. Other subclasses of nouns might deviate in some aspects from this definition, which is why they will be called "non-prototypical nouns".

Prototypical nouns have an inherently determined gender. Abesabesi's current gender system only distinguishes human from non-human, but it has probably developed from a noun class system with vowel-prefixes (see Section 3.3.2.2). They can also inflect for number, distinguishing singular from plural (see Section 3.3.2.1). A prototypical noun can head an NP and be modified by other nominals or by bound possessive pronouns. It carries a vowel harmony value, which is the source for harmonizing possessive pronouns. As head of an NP, it can function as a subject or object of a verb or as an adverbial adjunct.

In the first position of a sentence, it can carry the 1H, as a modified noun, it can carry the GenH or RelH when followed by a modifier, and as the subject of a question, it can carry the QH. If it denotes a location, it can be the complement of a preposition within a PP and carry the LocH.

Non-prototypical nouns are property nouns, numerals, function nouns, and deverbal nouns. All of them are discussed in detail in Section 3.3.3. Table 3.27 lists features of prototypical and non-prototypical nouns to distinguish the subclasses. Property nouns denote properties and usually modify nouns or pronouns. Numerals behave similar to property nouns, but denote numbers. They are analyzed as a subclass of nouns, as they cannot be formally distinguished from nouns. Function nouns express a grammatical or semantic relation towards another constituent of the clause. They express concepts that are expressed by prepositions in many other languages. Deverbal nouns are non-prototypical, as they only occur in limited positions in the clause. They are derived from verbs, but function syntactically as nouns.

Feature	Prototypical noun	Property noun	Numeral	Function noun	Deverbal nouns
Start with a vowel	+	+	+	+	+
Can be the head of an NP	+	+	+	+	+
Can be used as modifiers	+	+	+	–	+/–
Require GenH on head NP to be modifier	+	–	–	n.a.	+
Denote a property	–	+	+	–	–
Denote a numeral quantity	–	–	+	–	–
Express a grammatical or semantic relation towards another constituent of the clause	–	–	–	+	–
Derived from verbs	–	–	–	–	+

Table 3.27: Features of prototypical and non-prototypical nouns

3.3.2 Nominal morphology

Nouns in Abesabesi have a minimal word structure of V.CV. This structure can be extended by further CV syllables and/or a final coda consonant (see Section 3.1.3 for a detailed discussion on the syllable structure of nouns). Internal coda consonants only exist in loan-words. Thus, all nouns start with a vowel, which is indicative of a former noun-class system (see Section 3.3.2.2). This initial vowel must not be /u/ and vowel-harmony constraints apply to the vowels of all non-compounded, non-borrowed words. Borrowed nouns without an initial vowel are re-analyzed with an initial /i/, which does not have to be realized in isolation, but appears when preceded by another word.

Only limited morphology can be added to Abesabesi nouns. Apart from the noun-class prefix, nouns can take possessive pronoun suffixes and the goal-suffix *-é*.

3.3.2.1 Number

Abesabesi features number marking not only in the pronoun paradigms, but also on nouns, albeit the use of plural on nouns has decreased. Table 3.28 displays some singular and plural nouns. Abesabesi distinguishes the inherent human/non-human category in nouns, which is also reflected in plural marking. Human nouns switch their initial vowel to an /a/, while non-human nouns switch their initial vowel to an /i/.

Human			Non-human		
Gloss	Singular	Plural	Gloss	Singular	Plural
person	ɔni	ani	animal	ɛnam	inam
wife	oɲo	aɲo	leaf	afa	ifa
husband	owos	awos	cow	ɛna	ina
friend	ɔʃamis	aʃamis	hand	ɔbɔ	ibɔ

Table 3.28: Plural marking on nouns

Plural marking on humans is regularly used but seems to be optional, as there are a few cases where speakers used the singular although referring to multiple people. If used with a numeral higher than one, the plural marking does not necessarily become redundant in Abesabesi (see Example 3.44). Two irregular plurals are in use: *ìkpàr* as plural for *ɔban* 'children', and *ìkadáani* as plural for *ɔ̀kɔdɔ́ɔni* 'old person'.

(3.44) *awos* *mi* *ʃàj* *ídiàn* *ìdɛ́n*
man\PL REL be_short two those
'those two short men' (ibe121-00.155)

Plural marking on non-humans, however, is rarely used and labeled as archaic by speakers. Plural marking on certain non-human nouns is even judged incorrect by younger speakers.

Demonstratives, the determiner *ɛjɔ̀g* 'another', and property nouns can be marked for plural. If the head noun is marked for plural, they agree in number. Tables 3.29 and 3.30

show that their initial vowel is switched to an /i/ except for the proximal demonstrative *èéni*, which has a separate plural form (*ìdín*).

Singular	Plural
ɔhunɛ èéni 'this tree'	*ɔhunɛ ìdín* 'these trees'
ɔhunɛ èdɛ́n 'that tree'	*ɔhunɛ ìdɛ́n* 'those trees'
ɔhunɛ ɛjɔ̀g 'another tree'	*ɔhunɛ ijɔ̀g* 'other trees'
ɔhunɛ ɛdug 'a big tree'	*ɔhunɛ idug* 'big trees'
ɔhunɛ ɛŋa 'a new tree'	*ɔhunɛ iŋa* 'new trees'

Table 3.29: Number agreement with non-human nouns

Singular	Plural
ɔni èéni 'this person'	*ani ìdín* 'these people'
ɔni èdɛ́n 'that person'	*ani ìdɛ́n* 'those people '
ɔni ɛjɔ̀g (also *ɔ̀gɔni*) 'another person'	*ani ijɔ̀g* 'other people'
ɔni ɛdug 'a big person'	*ani idug* 'big people'
ɔni ɛŋa 'a new person (stranger)'	*ani iŋa* 'new people (strangers)'

Table 3.30: Number agreement with human nouns

Nominal modifiers usually do not agree in gender (human/non-human) with the head noun. Only the property noun for 'dead' has a human form (*ohuhu*) as opposed to the non-human form *ehuhu* (see Example 3.45)

(3.45) (a) *ìkokò ehuhu*
chicken dead
'a dead chicken' (ibe292-00.045)

(b) *ɔni ohuhu*
person dead
'a dead person' (ibe001-01.120)

3.3.2.2 Noun class

While only some languages of the region still have a productive noun class system (see Salffner (2009, pp. 59-61) for Ukaan), most languages only exhibit traces, such as the initial vowels in nouns. This is also the case in Abesabesi. These initial vowels, former noun class prefixes, can represent an affiliation to a semantic domain, mark number, and give a distinction between human and non-human. The former noun-class system is clearly decaying, as most of these traits have become optional. Assuming the quality of the initial vowel represented the affiliation to a certain noun class, as it is the case in neighboring Ukaan, some semantic domains can be spotted. Table 3.31 shows that most lexical items denoting people start with an /o/ or /ɔ/, most animals and all properties start with an /e/ or /ɛ/, and gerunds all start with an /i/.

These former noun classes are also reflected in the human/non-human distinction in the pronoun paradigms of Abesabesi. While the bound 3SG pronoun for humans is *ɔ́*,

o-/ɔ-	e-/ɛ-	i-
ɔni ‘person’	*ɛnam* ‘animal/meat’	*idàsaǹ* ‘going’
oɲo ‘wife’	*ehu* ‘rat’	*ibùgoǹ* ‘beating’
owos ‘man/husband’	*ɛbuj* ‘sheep’	*ikòoǹ* ‘falling’
otu ‘thief’	*ebo* ‘dog’	*iwùʃoǹ* ‘planting’
ɔmis ‘king’	*ɛna* ‘cow’	etc.
òtòj ‘girl’	*ɛdug* ‘big’	
ɔʃamis ‘friend’	*ɛmiɲɛ* ‘small’	
ɔ̀kɔdɔ́ɔni ‘old person’	*ɛŋa* ‘new’	
ɔban ‘child’	*ehĩs* ‘empty’	
oso ‘rival’	*ekuku* ‘old’	

Table 3.31: Initial vowels matched with semantic domains

the pronoun for non-humans is *ɛ́*. This distinction is also present in the 3PL pronouns *bà/bè/bò* (human) and *í* (non-human).

Number marking is another trace of the noun-class system. As discussed in the previous section, human plural nouns start with an /a/, while non-human plural nouns start with an /i/. Demonstratives and property nouns both agree in number with the noun they modify. There is only one instance of a property noun agreeing with a human noun by switching the initial vowel to /o/ (see Example 3.45).

Synchronically, Abesabesi’s system has reduced to four classes, distinguishing human from non-human and singular from plural (see Table 3.32). Agreement has been reduced to a singular/plural distinction, while pronoun paradigms distinguish all four classes.

Class	Prefixes	Agreement	Subject pronoun
Human Singular	mostly O-	E-	ɔ́
Human Plural	mostly A-	I-	bà/bè/bò
Non-human Singular	A-/E-/I-/O-	E-	ɛ́
Non-human Plural	I-/no vowel change	I-	í

Table 3.32: Synchronous analysis of noun classes

If we look at Ukaan, to compare Abesabesi to a productive noun class system of the region, we can see many similarities amongst the noun class prefixes (see Figure 3.10). For example, the prefix for the first noun class including mostly humans is also O- and its plural is A-. The other plural class has the prefix -I. In Abesabesi and Ukaan, the ATR value is not critical for the affiliation to noun classes, which is why capital letters are used as archiphones. Ukaan’s noun class system could be a possible representation of what Abesabesi’s system could have looked like before it lost the agreement classes.

This system is also similar to the Edoid one. Elugbe (1989, p. 123) reconstructs the five vocalic singular noun prefixes U-, I-, E-, O-, and A- and two plural prefixes A- and I-.

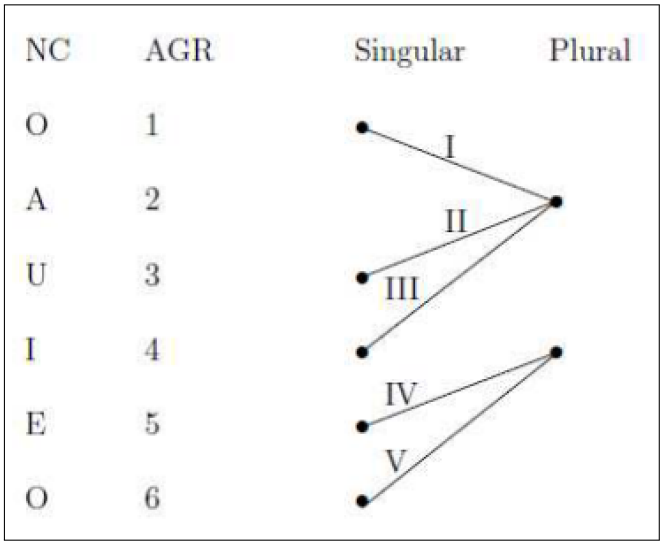


Figure 3.10: Ukaan noun classes. (Salffner, 2009, p. 60)

3.3.2.3 Possessive pronouns

Nouns can take pronoun suffixes that mark possession. The pronoun paradigm and its function is discussed in Section 3.5.2.3.

3.3.2.4 Goal

Abesabesi exhibits a suffix -é after nouns ending in a consonant and a H on the last vowel of nouns ending in a vowel. It is assumed that the H is a result of V_2 deletion, when -é meets a final vowel. This morpheme is only attached to nouns denoting a location and predominantly when filling the thematic role of a goal. For simplicity reasons, this category will be called the goal-suffix and glossed 'GOAL', but as it also appears in locative phrases, its exact semantics are yet to be investigated. The following examples should give an overview of its distribution.

(3.46) (a) *ɔ́ ke niŋ iʃo é → [ɔ́keniŋ iʃó]*
3SG PRF reach house GOAL
's/he has arrived at the house' (ibe332-00.117)

(b) *ǹ á das èʃî é → [ǹjádas èʃˡé]*
1SG PROG go market GOAL
'I am going to the market' (ibe265-00.190)

(c) *ǹ á das èkiròm é → [ǹjádas èkiròmˡé]*
1SG PROG go Ikaram GOAL
'I am going to Ikaram' (ibe346-00.130)

(d) *àti* *iniŋ* *mi* *á* *ʃɛlɛ̀* *H* *ègur* *é*
and thing REL PROG LOC town GOAL

→ [àtiniŋ íɲáʃɛlɛ̀ égùre]
'and the things that are happening in the town' (ibe096-00.008)

3.3.3 Non-prototypical nouns

A few word classes are analyzed as nouns, but they may not share all nominal features with prototypical nouns. These non-prototypical sub-classes of nouns are property nouns (Section 3.3.3.1), numerals (Section 3.3.3.2), function nouns (Section 3.3.3.3), and deverbal nouns (Section 3.3.3.4).

3.3.3.1 Property nouns

Abesabesi assigns properties through verbs (see Section 3.4.4.1) and nouns. The small closed word class of property nouns, in isolation, are not distinguishable from other nouns. Yet, they all denote a property (dimension, physical property, color, human propensity, age, or value) and start with a mid front vowel (/e/ or /ɛ/). Here follows an exhaustive list of all property nouns found in the corpus:

- *ɛdug* 'big'
- *ɛmiɲɛ* 'small'
- *ehij* 'old'
- *ètòj* 'young'
- *ekuku* 'old'
- *ɛŋa* 'new'
- *etito* 'hot'
- *elilоŋ* 'cold/wet/kind'
- *ebìbe* 'cooked'
- *ebisìŋ* 'raw'
- *ehĩs* 'empty'
- *ehuhu/ohuhu* 'dry/dead'
- *efo* 'white'
- *ehidino* 'black'
- *ɛŋa* 'red'

- *èɲèm* ‘good’
- *ɛʃa* ‘bad’

Like other nouns, property nouns can be the head of a noun phrase (see Example 3.47). This is mostly the case in sentences with the copula verb *tu*.

(3.47) (a) *ɛmiɲɛ tu èɲeɲèn*
small be yours
‘the small one is yours’ (ibe292-00.378)

(b) *ɛdug gí tu èɲeɲès*
big FOC be ours
‘the big one is ours’ (ibe292-00.377)

(c) *ɛsagum tu ehidino è*
sheep be black NEG
‘the sheep is not black’ (ibe365-00.179)

Property nouns in predicative use are often reduplicated (see Example 3.48). The morphological process of reduplication will be explained in Section 3.6.4.

(3.48) (a) *ìwé màá tu ɛŋaŋa*
book HAB be new
‘the book used to be new’ (ibe363-00.159)

(b) *iʃo tu efofo è*
house be white NEG
‘the house is not white’ (ibe365-00.178)

When used to modify other nouns, property nouns occur in attributive constructions (see Section 3.3.4). This construction consists of the modified noun followed by the property noun. In contrast to the genitive construction, the attributive construction does not feature a GenH.

Most property nouns do not appear very frequently, as they have verbal equivalents (see Section 3.4.4.1). These property verbs are used in relative clauses to modify nouns (see Example 3.49).

(3.49) (a) *òtùtù eliloŋ*
beer cold
‘a cold beer’ (ibe147-00.015)

(b) *bíà mi lelìŋ*
beer REL be_cold
‘a cold beer’ (ibe292-00.392)

Numeral	Meaning
ekìn/ígbon	one
ídiàn	two
ísas	three
íniŋ	four
íʃoǹ	five
íʃánas	six
íʃenèʃ	seven
ànaàniŋ	eight
òkpolòʃ	nine
ìjof	ten
ɔgbɔlɔ	twenty
ijum	four hundred

Table 3.33: Base numerals

3.3.3.2 Numerals

Numerals behave similar to property nouns, but denote a numeral quantity. There are twelve base forms in Abesabesi (see Table 3.33).

The rest of the numbers are composed in a vigesimal fashion except for the five numbers before 20 and its multiples (15 – 19, 35 – 39, etc). This means 11-14 are composed as x + 10; 21-34, 41-54, 61-74 etc. are composed as x + (20 × y); and 15-19, 35-39, 53-59, etc. are composed as x from (20 × y) (see Table 3.34)

As all the higher numbers have only been recorded once by one speaker, the transcription might not be accurate. For other accounts of the composition of numerals, consult Agoyi (2012c) or Oyebade (2015).

All these numbers are cardinal numbers used for counting and to modify nominals. The only exception is the number "one", where *ígbon* is only used for counting and *ekìn* is only used to modify nominals.

When modifying nominals, they behave similar to property nouns (Section 3.3.3.1). They follow the nominal they modify and do not trigger a GenH on the modified word. Within the corpus, numerals only modify nouns or the 3PL.HUM.IDP pronoun *àba*.

(3.50) (a) *ǹ ʃi ɔban ekìn*
1SG have child one
'I have one child' (ibe015-00.157)

(b) *àba ídiàn gí tu òden ò*
3PL.HUM.IDP two FOC be here EMPH
'two of them are here' (ibe009-00.038)

Numeral	Composition	Meaning
ekìntef	1 + 10	11
ídiàntef	2 + 10	12
ísaàtef	3 + 10	13
íniŋtef	4 + 10	14
íʃoǹlɔgbɔ̀lɔ̀/íʃoǹle	?	15
íniŋdɔgbɔ̀lɔ̀	4 from 20	16
ísáasdɔgbɔ̀lɔ̀	3 from 20	17
ídiàndɔgbɔ̀lɔ̀	2 from 20	18
ekìndɔgbɔ̀lɔ̀	1 from 20	19
ɔgbɔlɔ	20	20
ekinnɔ̀gbɔ̀lɔ̀	1 + 20	21
ídiànɔgbɔ̀lɔ̀	2 + 20	22
ísasínɔ̀gbɔ̀lɔ̀	3+ 20	23
íniŋínɔgbɔ̀lɔ̀	4 + 20	24
íʃonnɔgbɔ̀lɔ̀	5 + 20	25
íʃánasínɔgbɔlɔ	6 + 20	26
íʃenèʃínɔ̀gbɔ̀lɔ̀	7 + 20	27
ànaàniŋínɔ̀gbɔ̀lɔ̀	8 + 20	28
òkpolòʃńɔ̀gbɔ̀lɔ̀	9 + 20	29
ìjofńɔ̀gbɔ̀lɔ̀	10 + 20	30
ekìntefínɔ̀gbɔ̀lɔ̀	1 + 10 + 20	31
ídiàntefńɔ̀gbɔ̀lɔ̀	2 + 10 + 20	32
ísatefńɔ̀gbɔ̀lɔ̀	3 + 10 + 20	33
íniŋtefńɔ̀gbɔ̀lɔ̀	4 + 10 + 20	34
íʃonlénɔ̀gbɔ̀lɔ̀	15 + 20	35
íniŋídígbɔ́diàn	4 from (20 × 2)	36
ísasídígbɔ́diàn	3 from (20 × 2)	37
ídiàndígbɔ́diàn	2 from (20 × 2)	38
ekìndígbɔ́diàn	1 from (20 × 2)	39
igbɔdiàn	20 × 2	40

Table 3.34: Composed numerals

Ordinal numbers are formed by the word *èjà* 'number' followed by one of the cardinal numbers: E.g., *èjà ísas* 'third' or *èjà íʃenèʃ* 'seventh'. As *èjà* is a noun, the combination can be used to modify nouns (Example 3.51 a) or as the head of an NP (Example 3.51 b). For the ordinal number 'first', *ekìn* is not used but the Yoruba loan *àkɔ̀kɔ́*: *èjà àkɔ̀kɔ́* 'first'

(3.51) (a) *adɔm ɛ̀jà íʃenèʃ gí asaá màá munà*
month number seven FOC 1PL.HAB HAB celebrate_new_yam
'we celebrate the new yam in July' (ibe035-00.030)

(b) *ɛ́ bá wa ɛ̀jà ísas*
3SG.NHUM FUT be number three
'it must be the third one' (ibe111-00.082)

Distributive numbers are formed by full reduplication: *ìjof ìjof* 'ten each', *ekìn ekìn* 'one each/one by one' (Section 3.6.4).

3.3.3.3 Function nouns

This sub-class of nouns has been grouped together as they all have less of a semantic and more of a grammatical function and occur in genitive constructions with a possessor noun. The possessor noun provides the semantic information while the function noun modifies it grammatically. Most of these nouns express a spatial relation towards the possessor noun and are thus translated as spatial prepositions. They are usually derived from words for body parts. The noun *ìwùd* for example usually means 'stomach' but can be used as a function noun meaning 'inside' (Example 3.52). A function noun does not head a locative phrase or replace the LocH. If used as a location, it needs to carry the LocH.

(3.52) *ɔ́ tu H ìwùd iʃo*
3SG be LOC inside house
's/he is inside the house' (ibe121-00.286)

Table 3.35 lists all function nouns found in the corpus. Some of them have regular noun equivalents whose meanings are indicated as "original meaning". The others either don't have a regular noun equivalent or it is not known.

Some function nouns have functions apart from expressing a spatial relation. *ɔ̀kú* and *ɔ̀kpɔ̀lɔkpɔ̀* function as quantifiers. They act as heads in a NP and can be modified by nouns (Example 3.53 a) or possessive pronouns (Example 3.53 b). The dependent noun or possessive pronoun is the entity to be quantified.

(3.53) (a) *ɔ̀kpɔ̀lɔkpɔ̀ ani*
many people
'many people' (ibe265-00.085)

(b) *ɔ̀kú i*
all 3PL.NHUM.POSS
'all of them' (ibe301-00.073)

Function nouns do not have to be modified by a noun but can also take a possessive pronoun as a possessor.

Noun	Original meaning	Functional meaning
àgòr	buttocks	at the bottom of
asas	middle	between/among
èdʒì	?	like
ègido	?	only
ìbàbà	side	next to
ìbùg	?	from/among
ijes	body	against/PAR/RECP
iʃum	head	on top of/above
íʃa	?	for
ìwùd	stomach	in(side)
òkú	?	all
onu	mouth	around
orajo	?	front
òʃum	back	behind
ɔbɔ	hand	from
ɔ̀kpɔ̀lɔkpɔ̀	?	many

Table 3.35: Function nouns

(3.54) (a) *ɔ́ kèd ɛ kiʃ lí iʃum*
3SG.HUM carry 3SG.NHUM.OBJ back LOC on_top_of
ɛ
3SG.NHUM.POSS
's/he put it back on it' (ibe203-00.263)

(b) *ɔ́ mɔ̀ŋɔ̀ ɛmũ ɔbɔ na*
3SG.HUM borrow money from 1SG.POSS
's/he borrowed money from me' (ibe331-00.477)

Only three words *èdʒì* 'like', *ègido* 'only', and *íʃa* 'for' are different as they do not take possessive pronouns but independent pronouns (see Example 3.123 in Section 3.5.2.4). They could thus be considered prepositions, but as they have a nominal form and differ from the only other prepositions *lí* and *ʃí*, they are classified as functional nouns here.

3.3.3.4 Deverbal nouns

Nouns can be derived from verbs. Abesabesi has two strategies to create deverbal nouns: simple nominalization and complex nominalization. Both strategies are discussed in Section 3.6.1.

While the abstract verbal nouns created through simple nominalization are functionally equal to regular nouns, gerunds created through complex nominalization only occur

in a special construction - the relativized focus construction. Only a few gerunds are lexicalized and can be used like regular nouns: *imùnaǹ* 'new yam festival' from *munà* 'celebrate the new yam' or *ìguj̀ iwùʃoǹ* 'seed tubers' from *ìguj̀* 'materials' and *wuʃò* 'to plant'.

3.3.4 Noun phrase

Noun phrases are complex syntactic units that can be headed by any kind of noun or pronoun. In Abesabesi, they can function as arguments of a VP, adjuncts, or as dependents of other NPs or PPs. As arguments, NPs can be used as a subject, a predicate nominal, a primary object, or a secondary object. Therefore, they can occupy any of the three argument positions (before a verb, after a verb, or after the primary object), the adjunct position (at the end of a clause), the focus-fronted position, or the right- or left-dislocated positions.

An NP minimally contains a head. The head can be a noun or pronoun. Example 3.55 shows different examples of NP heads: A prototypical noun (3.55 a), a property noun and an independent possessive pronoun (3.55 b), a function noun (3.55 c), an independent pronoun (3.55 d), an indefinite pronoun (3.55 e), and a demonstrative pronoun (3.55 f). Optionally, the head can be modified by one or multiple dependents. Dependents can be other nouns, numerals, demonstratives, pronouns, or relative clauses.

(3.55) (a) *ida é dʒe*
fire PROG burn
'the fire is burning' (ibe012-00.066)

(b) *ɛmiɲɛ tu èɲeɲèn*
small be yours
'the small one is yours' (ibe292-00.377)

(c) *ɛ́ deŋ iʃum ísaj*
3SG.NHUM climb on rock
'it climbed onto a rock' (ibe346-00.072)

(d) *àbèn bɔ̀r atɔ*
2PL.IDP hoe ground
'you hoed the ground' (ibe128-00.015)

(e) *néè je ɔ́ntɔn è*
1SG.NEG see anybody NEG
'I did not see anybody' (ibe265-00.039)

(f) *èéni ga*
this be_enough
'this is enough' (ibe182-00.117)

There are three different constructions to connect a noun to a modifier: The attributive construction, the genitive construction, and the relative construction. Which kind of construction is used depends on the kind of modifier. Possessors (nouns or pronouns)

are connected with the genitive construction. Therefore, the head noun carries the GenH (Section 3.1.5.2) and is followed by the possessor. The relative construction connects relative clauses. Therefore, the head noun carries the RelH (Section 3.1.5.2) and is followed by the relative clause. In all other cases, the attributive construction is used, where the noun and the modifier are simply juxtaposed. A genitive or attributive construction only occurs between the modified noun and the first modifier. If more modifiers exist, they simply follow the first modifier.

The following subsections discuss each type of modifier and how they are used in an NP. The last subsection deals with the word order within an NP.

3.3.4.1 Nouns as modifiers

Nouns can be modified by other nouns. The modifying noun, as any other modifier, follows the modified noun. Nouns usually modify other nouns in a genitive construction, where they act as possessors. Property nouns and numerals, however, are connected through the attributive construction. Example 3.56 shows a noun modified by a prototypical noun, a property noun, and a numeral.

(3.56) (a) *iʃo* *H* *ɛna*
house GenH cow
'the cow's house' (ibe029-00.146)

(b) *iʃo* *ɛdug*
house big
'a big house' (ibe203-00.049)

(c) *iʃo* *ekìn*
house one
'one house' (ibe005-00.140)

3.3.4.2 Determiners

The four demonstratives and the determiner *ɛjɔ̀g* – 'another' (Section 3.5.1) can be used independently as pronouns or dependents that modify nouns. If used as modifiers, they follow the noun they modify and do not require the GenH. Plurality can be marked on determiners. If the head noun is marked for plural, determiners agree in number (see Section 3.3.2.1). Example 3.57 shows the determiner *ɛjɔ̀g* 'another' and its plural form *ijɔ̀g* modifying a noun.

(3.57) (a) *ɔni* *ɛjɔ̀g*
person other
'another person' (ibe140-00.126)

(b) *ani* *ijɔ̀g*
people other\PL
'those people' (ibe140-00.010)

3.3.4.3 Pronouns

The following pronouns can be used as modifiers in an NP: Possessive pronouns, indefinite pronouns, and interrogative pronouns. Like all nominal modifiers, they follow the head. Possessive pronouns, however, are suffixed and harmonize with the head's root (see Section 3.5.2.3). All pronouns are connected through the GenH.

(3.58) (a) *ɔban H na das iʃáafa*
child GenH 1SG.POSS go school
'my child went to school' (ibe203-00.245)

(b) *ɔban èéni tu ɔban H ɔ́ntɔn è*
child this be child GenH nobody NEG
'this child is nobody's child' (ibe265-00.046)

(c) *ɔban H máàn ɔ̀ tu*
child GenH who 2SG.SBJ be
'whose child is this?' (ibe017-00.032)

3.3.4.4 Relative clauses

The structure of relative clauses is discussed in Section 3.7.4.2. They follow the noun they modify and are connected through the RelH.

(3.59) *ɔ́ dʒe iniŋ edʒe H mi ʃen è*
3SG.HUM eat food RelH REL be_good NEG
's/he ate bad food' (ibe182-00.313)

3.3.4.5 Word order

The minimal NP consists of only the head. All modifiers follow the head. The quantifiers *ɔ̀kú* 'all' and *ɔ̀kpɔ̀lɔkpɔ̀* 'many' (a Yoruba loan-word) are not considered modifiers. They are nouns in their own right and form their own NP with a dependent noun or pronoun (Example 3.53). These two quantifiers, along with other function nouns, are discussed in Section 3.3.3.3.

The most complex NP elicited was the following:

(3.60) *N PRO NUM PN [ ]$_{REL}$ DET*
iʃo no ísas iŋa mi ǹ ta ìdín
house 1SG.POSS three new REL 1SG.S build these
'my three new houses that I built' (ibe292-00.420)

This example also displays the default order found in most NPs:
Head > PRO > PN/NUM > REL > DET

(3.61) (a) *N PN [ ]$_{REL}$*
wol màlúù ɛdug mi é dʒiniŋ
kill cow big REL PROG eat
'kill the big cow that is eating' (ibe292-00.400)

(b)

		N	*PN*	*NUM*	*DET*
ɔ́ˈɔ́	je	ìkokò	ifo	ídiàn	ìdɛ́n?
2SG.Q	see	chicken	white	two	those

'have you seen the two white chickens?' (ibe292-00.427)

(c)

N	*PRO*	*PN*	*[*		*]$_{REL}$*
ɔban	na	ɛmiɲɛ	mi	ʃàj	fáŋ
child	1SG.POSS	small	REL	be_young	SUP

'my youngest small child' (ibe292-00.597)

However, the order within the NP is not very rigid. Divergences from it are judged acceptable and infrequently occur in the corpus.

(3.62) (a)

			N	*[*	*]$_{REL}$*	*PN*
ǹ	báa	ŋɔ	bíà	mi	lelìŋ	ɛmiɲɛ ɛmiɲɛ
1SG	DES	beer	REL	be_cold	small	

'I want to drink a small cold beer' (ibe292-00.392)

(b)

		N	*PRO*	*DET*	*NUM*	*PN*
ɔ́ˈɔ́	je	iʃo	no	ìdín	ísas	iŋa?
2SG.Q	see	house	1SG.POSS	these	three	new

'have you seen my three new houses?' (ibe292-00.423)

(c)

N	*[*	*]$_{REL}$*	*NUM*	*DET*
awos	mi	ʃàj	ídiàn	ìdɛ́n
men	REL	be_short	two	those

'those two short men' (ibe121-00.155)

3.4 Verbs and the verb phrase

This section discusses the verb and its features (Section 3.4.1), the morphology of verbs (Section 3.4.2), TAM categories (Section 3.4.3), non-prototypical subclasses of verbs (Section 3.4.4), and the verb phrase 3.4.5.

3.4.1 The verb

The class of verbs is consonant-initial. Their main distinction from nominals is the ability to head a VP. The only exception are auxiliary verbs, which do not function as heads of VPs. However, they occupy verbal positions in the VP, start with a consonant and are able to take subject, TAM, and object affixes. This is why they are classified as verbs.

In order to classify verbs, a bundle of features will be introduced which define a prototypical verb. The non-prototypical subclasses of verbs have most of these features as well, but might lack one or a few of them.

The prototypical verb in Abesabesi opens one to three argument positions, which can be occupied by a subject, a primary object, and a secondary object. It takes bound subject pronouns, TAM markers, and bound object pronouns, which may harmonize with

Feature	Prototypical verb	Property verb	Auxiliary verb	Defunct verb
Start with a consonant	+	+	+	+
Can be main verb of a clause	+	+	–	+
Takes bound pronouns	+	+	+	–
Takes TAM markers	+	+	+	+/–
Can be transitive	+	–	+	–
Denotes a property	–	+	–	–

Table 3.36: Verbal subclasses

the verb's VH value. Moreover, it carries a GenH before object pronouns and triggers the 1H on the first-position element of the clause in affirmative clauses. Lastly, the gerund circumfix and the pluractionalis suffix may attach to a prototypical verb.

Non-prototypical verbs are property verbs, defunct verbs, and auxiliary verbs. They are each discussed in detail in Section 3.4.4. Table 3.36 lists features of prototypical and non-prototypical verbs. Property verbs denote properties and are the verbal equivalent of property nouns. They can be the main predicate of a clause or modify nominals as part of a relative clause. Defunct verbs do not trigger the 1H and only take one argument. Auxiliary verbs do not provide the main semantic information of a VP but modify the verb. They either add a participant to the event or modify the main verb in terms of time, direction, aspect, manner, cause, or comment.

3.4.2 Verbal morphology

Verbs, in comparison to nouns, always start with a consonant. The minimal verb has the structure CV, which can be extended by further CV syllables and/or a final coda consonant. The maximum verb in this database is CV.CV.CV. Additional verbal morphology will be discussed in the following subsections. Affixes can occupy one of two prefix and two suffix slots. The first prefix slot is reserved for bound subject prefixes, the second for TAM prefixes. The first suffix slot is reserved for the pluractionalis extension, and the second for object suffixes. Nominalization affixes are discussed in the section about word formation (3.6).

$$Subject\ pronoun - TAM - Verb - Pluractionalis - Object\ pronoun$$

3.4.2.1 Bound subject pronouns

The first prefix slot is reserved for bound subject pronouns. They are used to substitute NPs in subject position. All bound subject pronouns and their distribution are discussed in section 3.5.2.1.

3.4.2.2 Tense aspect mood markers

Tense aspect mood (TAM) markers are a class of prefixes that attach to verbs. Their position is between bound subject pronouns and verbs. Some of them can be combined, but the semantics of combined TAM markers is not completely understood and only briefly discussed in Section 3.4.3.9.

Abesabesi features nine distinct TAM markers: the progressive marker *á/é/ó*, the future marker *bá/bé/bó*, the desiderative marker *báa/bée/bóo*, the perfect marker *ka/ke/ko*, the negative perfect marker *kinà/kinè/kinò*, the habitual marker *màá/mèé/mòó*, the sequential marker *si/sí*, the state-changing marker *sì*, and the epistemic necessity marker *bá wa/bé we/bó wo*. The lack of TAM markers will be referred to as the "factative" category. All markers, except the sequential marker *si/sí* and the state-changing marker *sì*, have three forms as they are subject to prefix-VH caused by the verb stem. A detailed discussion of their function is given in Section 3.4.3, where the TAM markers along with other devices to express TAM categories are presented.

3.4.2.3 Object pronouns

Substitutes for object NPs, if not focus fronted as independent pronouns, have to be expressed as object pronoun suffixes on the verb. They occupy the second suffix slot after the pluractionalis extension. The pronoun paradigm and its function is discussed in Section 3.5.2.2.

3.4.2.4 Pluractionalis

Abesabesi features one verbal extension – the pluractionalis. This verbal suffix has six different allomorphs, as it depends on the verb's length and is affected by dissimilation and VH. All forms obey to suffix-VH. [+ATR] verbs take allomorphs with /o/ vowels and [-ATR] verbs take allomorphs with /a/ vowels. Verbs with a CV structure take the long allomorphs *-odiòg* or *-adiàg*. Longer verbs, on the other hand, take shorter allomorphs of the structure -VC. While the vowel adheres to VH, the final consonant is either /g/ or /d/. If the verb's last consonant is velar, the suffix's consonant is /d/, and otherwise, it is /g/. The six allomorphs are thus *-adiàg*, *-odiòg*, *-àd*, *-òd*, *-àg*, and *-òg*.

Verbs with the pluractionalis suffix lose their lexical tone and adopt a tonal melody of ML. This means all TBUs bear a mid tone except the last one, which bears a low tone.

The pluractionalis suffix is a morpheme that attaches to finite verbs but can also be found on nominalized verbs (Example 3.63). The term "pluractionalis" is used here in the

Allomorph	Verb	Meaning	Verb + Allomorph
-adiàg	jɛ̀	slice	jadiàg
-odiòg	de	buy	dodiòg
-àd	baŋ	tie	baŋàd
-òd	bug	beat	bugòd
-àg	faǹ	press	fanàg
-òg	bis	wipe	bisòg

Table 3.37: Pluractionalis allomorphs

sense of Newman (1990, pp. 53-54). It is a verbal category that highlights the plurality of a nominal argument and/or the multiplicity of an event (see Example 3.64). The suffix does not occur very frequently because it is not an obligatory agreement marker. A differentiated semantic analysis is not possible, as it occurs less than 15 times in non-elicited texts of the corpus.

(3.63) *ɔ̀-fanàg*
NMLZ-press\PLUR
'bread' (ibe190-00.124)

(3.64) (a) *ida mi bà lamàg ìdín*
light REL 3PL light\PLUR these
'these lights that they switch on' (ibe096-00.024)

(b) *bò bó bugòd ijes ʃí èʃì*
3PL FUT beat\PLUR RECP at market
'they will beat each other at the market' (ibe140-00.062)

(c) *owos èéni dodiòg ɔban u*
man this search\PLUR child 3SG.POSS
'this man searched for his child' (ibe224-00.143)

(d) *òdè màá lagàd*
stool HAB break\PLUR
'the stool used to break' (ibe363-00.154)

3.4.3 Tense, aspect and modality

Tense, aspect and modality categories are not completely understood. This section will provide an overview of the categories that I was able to isolate. Most TAM morphemes can be combined in order to develop a new temporal or aspectual meaning. These combined uses still need further investigation. A short discussion will be given in Subsection 3.4.3.9.

3.4.3.1 Factative

Abesabesi exhibits a basic temporal distinction of future and non-future. While future is marked with *bá/bé/bó*, the non-future has no marker. It encompasses the present and the past. The verbal category where the verb appears with no TAM marker will be called the "factative" in the sense of Welmers (1974, pp. 346-347). While unmarked dynamic verbs have a default past reading (Example 3.65 a), unmarked stative verbs (including all property verbs) have a default present reading (Example 3.65 b). Lexical time references, however, can alter the default temporal reading (see Example 3.65 c)

(3.65) (a) *bò* *∅* *dʒo* *bo* *èkiròm̀* *òhun*
3PL.HUM FAC give_birth 3PL.OBJ Ikaram here
'They gave birth to them here in Ikaram' (ibe065-00.002)

(b) *òli* *èéni* *∅* *lelìŋ*
cloth this FAC be_wet
'this cloth is wet' (ibe147-00.027)

(c) *ijɛ̀n,* *ɛ́* *∅* *lelìŋ*
yesterday 3SG.NHUM FAC be_cold
'yesterday, it was cold' (ibe178-00.190)

3.4.3.2 Future

The future tense is expressed through the marker *bá/bé/bó*. This tense sets an action or state into a temporal context that is yet to come. The marker harmonizes with the verb root.

(3.66) (a) *ǹ* *bé* *ko* *ohumò* *ekìn* *sá*
1SG FUT sing song one 2SG.DAT
'I will sing a song for you' (ibe140-00.454)

(b) *ɔ̀gɔni* *bá* *hɔ̃* *iʃòmò* *ɛjɔ̀g*
another_person FUT cultivate farm another
'another person will cultivate another farm' (ibe167-00.055)

(c) *ɔ́* *bó* *kùʃo*
3SG.HUM FUT come_down
's/he will come down' (ibe001-01.197)

3.4.3.3 Perfect

The perfect aspect refers to actions or events already completed before the reference time. It sets a focus on the resulting state after an action has happened. Because it combines with the basic tenses (future/non-future), it is analyzed as an aspect instead of a tense. The perfect is expressed through the marker *ka/ke/ko* in affirmative clauses (Example 3.67 a, b) and through *kinà/kinè/kinò* within a negative scope (Example 3.67 c, d). Both of them harmonize with the verb root. In almost all cases, the perfect aspect could be translated by 'already' (affirmative) or 'not yet' (negated).

(3.67) (a) *ɔ́ ko hu*
3SG.HUM PRF die
'he is dead' (ibe178-00.018)

(b) *ɔ́ˡɔ́ ka sɛmɛ̀ ohij so?*
2SG.Q PRF greet older_sibling 2SG.POSS
'have you already greeted your older sibling?' (ibe140-00.168)

(c) *aséè kinè dʒiniŋ è*
1PL.NEG PRF.NEG eat NEG
'we have not eaten' (ib017-00.153)

(d) *òdè kinà làg è*
stool PRF.NEG break NEG
'the stool is not broken' (ib363-00.128)

Stative verbs with perfect aspect express a state that has started or that is the result of a change. It is thus emphasized that the state had a starting point in the past.

(3.68) *ɔban èdén ka ga*
child that PRF be_big
'the child has become big/is already big' (ibe117-00.134)

The perfect aspect is barely used in the narrative context and can be combined with other aspect markers (such as the progressive). These criteria are the reason why this category is analyzed as perfect instead of perfective (Dahl, 1985, p 138).

3.4.3.4 Progressive

The progressive aspect refers to an ongoing action at the time of reference. It is expressed through the marker *á/é/ó*. The marker can be combined with the future tense marker or the perfect marker. Stative verbs are not used in the progressive aspect, as states are already continuous events.

(3.69) (a) *bà á ŋɔ iʃʷag*
3PL.HUM PROG drink palm_wine
'they are drinking palm wine' (ibe059-00.050)

(b) *ida é dʒe iʃo*
fire PROG burn house
'the house is burning' (ibe012-00.054)

(c) *mi à é tì ɛ*
SIP 1PL PROG push 3SG.NHUM.OBJ
'we were pushing through' (ibe065-00.120)

The progressive marker's basic form is just a high toned vowel which harmonizes with the verb root. If it is preceded by a nasal, such as the 1SG and 2PL realis pronouns, it is pronounced *já/jé/jó*.

(3.70) (a) *ǹ á ŋɔ iŋi* → *[ǹjáŋiŋi]*
1SG PROG drink water
'I am drinking water' (ibe005-00.209)

(b) *ida é dʒe iʃo* → *[idá édʒiʃo]*
fire PROG burn house
'the house is burning' (ibe012-00.054)

3.4.3.5 Habitual

The habitual aspect refers to recurring events. Verbs in the habitual aspect can either have a past or present reading. They can thus denote an event that frequently occurred in the past or one that has regularly occurred until the present and is expected to also occur in the future. Abesabesi marks the habitual aspect through the marker *màá/mèé/mòó*. It harmonizes with the verb root and requires the habitual pronoun paradigm (Section 3.5.2.1).

(3.71) (a) *ɔ́ màá kɔ ìwé*
3SG.HAB HAB write letter
'S/he usually writes letters/used to write letters' (ibe178-00.103)

(b) *eji màá kpɔ̀ hɔ àsìkò ɛ̀dɛ́n*
rain HAB be_much too_much time that
'there usually is too much rain at that time' (ibe167-00.152)

(c) *beé mèé je ɛmũ lí ɔbɔ*
3PL.HAB HAB see money LOC hand
'they usually have money to spend' (ibe140-00.070)

(d) *ɛɛ́ mèé lelìŋ*
3SG.NHUM HAB be_cold
'it usually is cold' (ibe178-01.184)

3.4.3.6 Sequential

The sequential marker *si/sí* marks the sequential aspect. It does not harmonize with the verb root and always comes first of all potential markers. The sequential aspect is used to mark the consecutiveness or simultaneity of a second, third, etc. event. While the first clause does not require the sequential marker, all following simultaneous or consecutive events are expressed as clauses with sequential marker. For instance, Example 3.72 contains four consecutive clauses where the first one is the matrix clause without sequential marker, and the following ones with sequential markers. Clauses with sequential marker do not have their own subject but always refer back to the matrix clause's subject. Thus, the sequential marker is never preceded by a bound subject pronoun.

(3.72) (a) *mi ɔ́ kùʃo ɛ̀dɛ́n*
SIP 3SG.HUM come_down that
'he came down from there' (ibe023-00.016)

(b) *sí* *be* *niŋ*
SEQU come arrive
'he arrived' (ibe023-00.016)

(c) *sí* *sùmùdò*
SEQU kneel
'he kneeled down' (ibe023-00.017)

(d) *sí* *á* *wuʃ* *i* *lí* *atɔ*
SEQU PROG put 3PL.NHUM LOC ground
'and put them on the ground' (ibe023-00.018)

3.4.3.7 Modality

Abesabesi exhibits a modal system consisting of TAM markers and auxiliary verbs to express modality. We distinguish the modal categories: imperative, desiderative, obligative, habilitative, and epistemic necessity.

Imperative The imperative is used to express commands, orders and requests. It distinguishes the second person singular from plural and affirmative from negative.

In affirmative imperatives of the second person singular, the subject is not overtly expressed. It consists of the unmarked verb and objects where required.

(3.73) *ʃiga* *imɔs*
rise sleep
'wake up!' (ibe140-00.154)

In all other cases, the subject has to be overtly expressed through a bound subject pronoun, but without the 1H. Negative imperatives require the irrealis pronouns.

(3.74) (a) *àn* *ɲadum*
2PL.SBJ sit
'sit down!' (ibe005-00.219)

(b) *séè* *ɲì* *èɡɛ̀* *ʃa* *ba* *è*
2SG.NEG open door DAT 3PL.OBJ NEG
'do not open the door for them' (ibe073-00.181)

(c) *anáà* *ŋɔ* *iŋi* *è*
2PL.NEG drink water NEG
'do not drink water!' (ibe070-00.067)

Desiderative The desiderative expresses intentions. In Abesabesi, it is marked by the TAM prefix *báa/bée/bóo*. It is similar to the future marker and likewise harmonizes with the verb root, but has a lengthened vowel with a HM melody.

(3.75) (a) *ǹ* *bée* *je* *dɔ́ktɔ̀*
1SG DES see doctor
'I want to see a doctor' (ibe073-00.195)

(b) *ǹ* *báa* *ŋɔ* *iŋi*
1SG DES drink water
'I want to drink water' (ibe104-00.055)

An alternative construction to express intentions or wishes involves the verb *dò* 'want' and a complement clause.

(3.76) *ǹ* *dò* *mí* *ǹ* *ŋɔ* *iŋi*
1SG want COMP 1SG drink water
'I want to drink water' (ibe017-00.124)

The desiderative prefix is also used to mark circumstantial necessity.

(3.77) *ǹ* *bée* *kpíʃ*
1SG DES sneeze
'I have to sneeze/I want to sneeze' (ibe111-00.046)

Obligative The obligative expresses an obligation. Abesabesi uses different variants of the Yoruba loan-auxiliary *gbɛ́dɔ̀* 'must' for strong deontic necessity and the construction *ɛ́ je ɛ mí* 'should/to be supposed to' (lit.: 'it sees it that...') for week deontic necessity.

(3.78) (a) *à* *gbúdɔ̀* *kpas* *àmùn* *èéni*
1PL must understand language this
'we have to understand this language' (ibe140-00.098)

(b) *ɛ́* *je* *ɛ* *mí* *ɔ̀* *ke* *niŋ* *òjè*
3SG.NHUM see 3SG.NHUM.OBJ COMP 2SG PRF arrive place
isum̀ *èhen*
work now
'you are supposed to be at your work place now' (ibe111-00.206)

Habilitative The habilitative expresses epistemic, deontic, and circumstantial possibility. Abesabesi uses the auxiliary *jɔbɔ* 'can/could' to express the habilitative.

(3.79) (a) *ògbegel* *èdɛ́n* *jɔbɔ* *á* *jas* *ani* *ɔ̀gbɔ̀lɔ̀*
car that can PROG take people twenty
'that car can take twenty people' (ibe117-00.104)

(b) *ɔ́* *jɔbɔ* *á* *ba* *ìdʒɛ*
3SG.HUM can PROG come today
's/he may be coming today' (ibe111-00.053)

(c) *ɔ́ɔ̀* *jɔbɔ* *das* *iʃáafa* *è*
3SG.NEG can go school NEG
's/he cannot go to school' (ibe140-00.389)

Epistemic necessity Epistemic necessity is usually expressed through the TAM markers *bá wa/bé we/bó wo*. Example 3.80 shows a situation where the fact that Ajoke is sneezing is only inferred from other information. The two TAM markers harmonize with the verb *kpiʃ* 'sneeze'.

(3.80) *àdʒɔkɛ́* *bé we* *kpiʃ*
<NAME> EPN sneeze
'Ajoke must be sneezing' (ibe111-00.048)

Example 3.81 also shows epistemic necessity. However, *bá wa* does not have a verb to harmonize with. Instead, *wa* seems to be the verb although the locative copula verb *tu* would be expected.

(3.81) *fűmi* *bá* *wa* *èʃì* *èhen*
<NAME> FUT be market now
'Funmi must be at the market now' (ibe111-00.028)

This indicates that the epistemic necessity marker *bá wa/bé we/bó wo* is in fact a combination of the future marker and the copula verb *wa*. The verb *wa* is otherwise used as a copula verb in hypothetic situations, such as counterfactual or hypothetical conditionals, equations in the future, or proper inclusions in the future (Section 3.7.2). When used to express epistemic necessity in a clause without copula verb, *wa* is used like a TAM marker and harmonizes with the verbal root (as in Example 3.80).

3.4.3.8 Other markers

The state-changing marker *sì* marks a change of state and can be translated as 'anymore' in negative sentences or 'now' in affirmative sentences. The marker does not strictly express a TAM category, but occupies a TAM slot. It cannot be analyzed as an auxiliary verb since it does not trigger its own VH but lets pronouns and other TAM markers to its left harmonize with the verb following *sì* (Example 3.82).

(3.82) (a) *náà* *sì* *tam* *iniŋ* *mi* *kpɛ̀* *è*
1SG.NEG SCM remember thing REL remain NEG
'I do not remember anymore what is missing' (ibe023-00.098)

(b) *néè* *sì* *je* *ɛ* *è*
1SG.NEG SCM see 3SG.NHUM.OBJ NEG
'I do not see it anymore.' (ibe096-00.041)

Another marker is *ba/be/bo*. This marker can carry a mid or low tone and is often found in narrative texts. Its semantics are not understood. It is simply glossed PST in examples.

(3.83) *mi* *èʃò* *bè* *kiȷ̃* *das* *òjè* *ɔni* *é* *mĩ* *olim̀*
SIP tortoise PST return go place person PROG make medicine
ò
EMPH
'tortoise returned to the herbalist's' (ibe111-00.048)

3.4.3.9 Combined TAM categories

As mentioned before, the presented TAM markers occur in combination with other TAM markers and seem to form new temporal or aspectual meanings. Many combinations are not understood, as some occur rarely and others are hard to separate into separate markers. Therefore, only some more frequently used combinations are presented here. An overview of the literature on TAM categories and their combinations is given below.

The progressive marker *á/é/ó* can often be found after the future marker *bá/bé/bó* (Example 3.84 a) or the perfect marker *ka/ke/ko* (Example 3.84 b, c). While the former expresses an ongoing action in the future, the latter expresses an action that was ongoing until the present or before another action started.

(3.84) (a) *bè* *bé* *é* *jo*
3PL.HUM FUT PROG dance
'they will be dancing' (ibe096-00.149)

(b) *ɔni* *ka* *á* *ʃɛ̀ǹ* *kpɔ̀* *hɔ*
person PRF PROG talk be_much too_much
'the person has been talking too much' (ibe117-00.143)

(c) *ɛ́* *ka* *á* *lagàd* *na*
3SG.NHUM PRF PROG itch 1SG.OBJ
'it has been itching me' (ibe331-00.614)

The combination of the perfect marker and the progressive marker with a stative verb expresses a development towards the state (Example 3.85).

(3.85) *òli* *no* *ka* *á* *lelìŋ*
cloth 1SG.POSS PRF PROG be_wet
'my clothes are getting wet' (ibe147-00.180)

Another common combination is the perfect marker *ka/ke/ko* followed by the future marker *bá/bé/bó*. It is frequently found in the matrix clause of counterfactual conditionals – the consequence to a condition that cannot or is unlikely to be fulfilled.

(3.86) *ɛ́ˡɛ́* *wa* *kpé* *ǹ* *sà* *kpé ní* *ɔ̀* *á* *ba,* *ǹ*
3SG.NHUM.COND be COMP 1SG know COMP 2SG PROG come 1SG
ka *bá* *kèd* *so* *èkó*
PRF FUT carry 2SG.OBJ <PLACE-NAME>
'if I had known you were coming, I would have picked you up in Lagos'
(ibe281-00.136-137)

Agoyi (2008, pp. 105-112) mentions TAM markers, as they are subject to vowel harmony. She analyzes them as preverbs but does not mention their function. Translated examples without glossing, however, give an idea of their use. Her examples match the simplex forms *bá/bé/bó* 'FUT', *ka/ke/ko* 'PRF', *á/é/ó* 'PROG' and the unmarked factative. Furthermore, she presents a few combinations of these simplex forms with translations:

TAM markers	Example	Translation
PRF + PROG	*ɔ́ ke é de*	'he is already buying'
FUT + PROG	*ɔ́ bé é de*	'he will be buying'
PRF + FUT	*ɔ́ ke bé de*	'he must have bought'
PRF + PROG + FUT	*ɔ́ ke é bé de*	'he might have bought'
PRF + FUT + FUT + PROG	*ɔ́ ke bé bé é de*	'he must have been buying'

Table 3.38: Combinations of TAM markers according to Agoyi (2008, p. 106)

In Agoyi (2012b, pp. 248-249), she analyzes VH in the Oshugu dialect and treats TAM markers in a separate section. Before Agoyi discusses the VH of Oshugu TAM markers in her article, she provides a sketch of Abesabesi's TAM system by presenting the following figure.

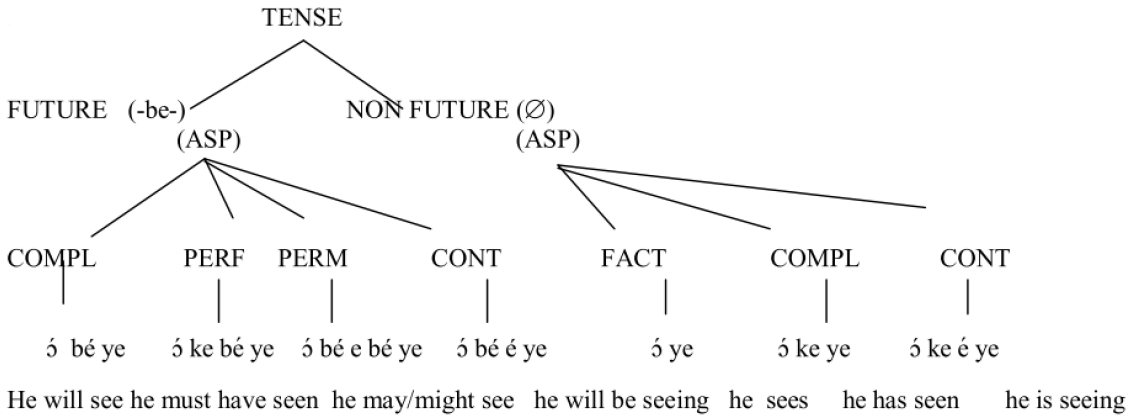


Figure 3.11: Abesabesi's TAM system according to Agoyi (2012b, p. 249)

She explains that Abesabesi distinguishes two tenses – future (marked with bV[10].) and non-future (unmarked) – and four aspects: completive, continuous, permissive, and perfective (Agoyi, 2012b, p. 248). The abbreviation "FACT" in the figure is not indicated, but might stand for factative. These categories' functions are not entirely clear to me, as no further explanations are given and the only clues are the category labels and the examples' translations.

3.4.4 Non-prototypical verbs

Besides the more prototypical verbs, Abesabesi exhibits subclasses of verbs that might deviate slightly from prototypical verbs, be it functionally or formally. The subclasses discussed in the following subsections are property verbs, auxiliary verbs, and defunct verbs.

[10]The vowel depends on the dialect's VH type

3.4.4.1 Property verbs

The subclass of property verbs contains static, intransitive verbs that can assign a property. The denoted properties include dimensions, physical properties, colors, human propensities, values, and quantities (see Examples in Table 3.39)

Property type	Verb	Meaning
Dimension	ʃàj	be small
	ɲidùg	be long
	dʒomò	be far
Physical property	honò	be hard
	ɲum̀	be sharp
	to	be hot
Color	ɲà	be red
	hidinò	be black
Human propensity	lelìŋ	be kind
	to	be wicked
Value	ʃen	be good
	wɔ̀f	be tasty
	ɲum̀	be sweet
Quantity	kpɔ́	be much/many

Table 3.39: Semantic domains of property verbs (non-exhaustive list)

Abesabesi can assign properties through property nouns (Section 3.3.3.1) and property verbs. The latter occurs more frequently and covers a wider range of properties. Moreover, property verbs are an open class, integrating even loan words such as Yoruba verbs *gɔ̀* 'be stupid' or *kpɔ̀* 'be much/many' and English adjectives *dɔ̀tí* 'be dirty'.

Property verbs assign a property to the subject of the clause or to any noun phrase through a relative clause containing the property verb (Example 3.87)

(3.87) (a) *òli* *èéni* *honò*
fabric this be_hard
'this fabric is stiff' (ibe292-00.210)

(b) *ísaj* *mi* *honò* *gí* *èéni*
stone REL be_hard FOC this
'this is a hard stone' (ibe292-00.219)

As property verbs are static verbs, they do not trigger a past tense interpretation when used in the zero-marked factative. Instead, they trigger a present tense interpretation (as seen in Example 3.88 a). Section 3.4.3 discusses the different behavior of active and static verbs in more detail. This behavior also distinguishes them from verbs expressing a process. Process verbs, such as *hij* 'become old' or *hoj* 'grow' are usually used with the PRF morpheme in order to express a property (the result of the denoted process).

(3.88) (a) *ìmúsù ɛjɔ̀g ∅ ga*
cat other FAC be_big
'the other cat is big' (ibe292-00.449)

(b) *ebo ke hij*
dog PRF become_old
'the dog is old/has become old' (ibe331-00.314)

3.4.4.2 Auxiliary verbs

Abesabesi has a few verbs that can be part of a verb phrase without constituting its head. They may modify the event expressed by the main verb of the verb phrase or add more thematic roles to it. As these verbs do not provide the main semantic information of a verb phrase but rather add additional information, I call them auxiliary verbs. Functionally, they can be grouped into participant-adding auxiliary verbs, which add a participant to the event, and adverbial auxiliary verbs, which modify the main verb by adding context (e.g., temporal, modal, manner, causal). While the participant-adding auxiliary *gbà* in Example 3.89 a adds the instrument role to the VP, the adverbial auxiliary *ʃisɛ̀n* in Example 3.89 b adds a temporal context.

(3.89) (a) *oŋo no gbà ɔbɔ bɛ̀s na*
wife 1SG.POSS put hand touch 1SG.OBJ
'my wife touched me with her hand' (ibe327-00.062)

(b) *bà ʃisɛ̀n ʃékinò ìwùd iʃo bo*
3PL just exit inside house 3PL.POSS
'they have just come out of their house' (ibe008-00.013)

Formally, they can be grouped into pre- and postverbs, as they appear in a specific position towards the main verb – either preverbal or postverbal. While *wɛ̀n* in Example 3.90 a is always situated before the main verb, *gɛn* in Example 3.90 b is always situated after the main verb.

(3.90) (a) *mí ǹ wɛ̀n niŋ*
COMP 1SG before arrive
'before I arrived' (ibe190-00.157)

(b) *ɔ́ˡɔ́ sù àlàj gɛn*
3SG.COND stick yam_stakes finish
'after s/he sticked the yam stakes' (ibe167-00.101)

Two adverbial auxiliary verbs occur before or after the verb: *tɛtɛ* 'slowly' and *jànà* 'early'. The latter only appears in complete reduplication after the verb. The position of these two auxiliaries does not change the meaning.

Some auxiliary verbs have an equivalent that can function as a full verb. The meaning of the auxiliary verb, however, is slightly different. Example 3.91 shows the full verb *kij*, which can mean 'return' when intransitive (a) and 'repeat' when transitive (b). The auxiliary verb *kij*, on the other hand, is used postverbally and means 'back' (Example 3.91 c).

(3.91) (a) *[...] ɔ́ kìj lí okpo ìdʒɔ̀*
3SG return LOC road farm
's/he returned to the farm road' (ibe140-00.421)

(b) *bè bé kìj ɛ kàs*
3SG FUT repeat 3SG.NHUM rake
'they will rake it again' (ibe035-00.006)

(c) *ɔ́ɔ̀ sɛmɛ̀ kìj è*
3SG.NEG greet back NEG
's/he did not greet back' (ibe265-00.348)

The verbal prefixes discussed in Section 3.4.2 may be situated before the preverb or the main verb with some restrictions. Bound subject pronouns always attach to the first verb: either the first auxiliary, if it is a preverb, or the main verb, if the auxiliary is a postverb. TAM markers have two possible slots. Either before the first verb (preverb or main verb) or between the preverb and the main verb. This second slot, however, may alter the meaning of the auxiliary. Example 3.92, for instance, shows that the auxiliary *ton* adds a beneficiary without the PROG marker *á/é/ó* in the second TAM slot (Example 3.92 a) and adds the role of a partner with the PROG marker in the second TAM slot (Example 3.92 b). TAM markers never appear between the main verb and a postverb. The following formula shows the order of preverbs, main verb, postverbs, and the TAM morphemes. Object positions are not included.

$$(PRO) - TAM_1 - Preverb(s) + TAM_2 - MainVerb + Postverb(s)$$

(3.92) (a) *ńkɔ̀ ton no sɛmɛ̀ ɔmis*
please meet 1SG.OBJ greet king
'please greet the king for me' (ibe039-00.018)

(b) *bà ton no á kɔn*
3PL meet 1SG.OBJ PROG fight
'they fight against me' (ibe057-00.159)

The TAM markers *ka/ke/ko* 'PRF' and *màá/mèé/mòó* 'HAB' usually only appear in the first TAM slot (an exception is the auxiliary *kɔ̃̀* 'just'), while *bá/bé/bó* 'FUT' and *á/é/ó* 'PROG' can occur in both slots.

As for VH, the main and auxiliary verb both open their own VH range. Bound subject pronouns and TAM markers always harmonize with the verb that follows directly after the TAM markers. Object pronouns harmonize with the verb that precedes. The fact that each auxiliary verb takes verbal affixes, such as bound subject pronouns, and triggers their VH is an indicator that auxiliaries are verbs and not a separate adverb word class. They also cannot be treated as verbal affixes themselves, as they do not harmonize, but trigger their own VH. For instance, Example 3.93 shows the subject pronouns and the PST marker harmonizing with the preverb *dʒɔ* and the PROG marker harmonizing with the main verb *mĩ*.

(3.93) *bà bà dʒɔ é mĩ isum̀*
3PL PST together PROG do work
'they worked together'

(ibe065-00.136)

The following distinction between participant-adding auxiliary verbs and adverbial auxiliary verbs is based on the verb's valency. Participant-adding verbs are transitive and add a participant to the event, while adverbial verbs are intransitive and modify the main verb. Two auxiliary verbs are part of both groups: *das* and *fàŋ*. They can be participant-adding or adverbial verbs depending on whether they come with an argument. *Das*, which has the meaning 'to go to' as a main verb (3.94 a), marks a goal as participant-adding verb 3.94 b) and signifies a movement away from the speaker as an adverbial verb (3.94 c)

(3.94) (a) *ǹ bá das àkùŋgbá*
1SG FUT go Akungba
'I will go to Akungba' (ibe017-00.292)

(b) *ìkpàr màá ʃu das ègur*
children HAB run go city
'children usually run to the city' (ibe096-00.003)

(c) *ɔ́ bɛ̀bɛ̀rɛ̀ á ʃu das*
3SG quickly PROG run go
's/he quickly ran away' (ibe140-00.233)

The verb *fàŋ* cannot function as the main verb. If used as a transitive auxiliary, it introduces a comparator argument (Example 3.95 a). When used as an adverbial verb, it marks the verb as superlative (Example 3.95 b).

(3.95) (a) *èéni ɲidùg fàŋ ɛ̀dɛ́n*
this be_long CPOR/SUP that
'this is longer than that' (ibe121-00.265)

(b) *ɛbuj èéni gí ga fàŋ*
goat this FOC be_big CPOR/SUP
'this goat is the biggest' (ibe147-00.153)

Participant-adding auxiliaries Abesabesi exhibits transitive auxiliaries that add a participant to the event. Some purely intransitive full verbs even rely on these auxiliaries in order to add arguments to the VP. Each auxiliary can introduce one thematic role to the event. Depending on the auxiliary, the role can be a patient, beneficiary, recipient, partner, comparator, instrument, goal, or source. Example 3.96 a, for instance, displays a simple clause with the intransitive verb *ba* 'come'. In order to add a source, the participant-adding auxiliary *wa* is used (Example 3.96 b). Likewise, a goal can be added to a clause with the intransitive verb *ʃiŋ̀* 'walk' (Example 3.97 a,b).

(3.96) (a) *ɔ́ ka ba*
3SG.HUM PRF come
's/he has come' (ibe121-00.184)

(b) *ɔ́ ka wa iʃáafa ba*
3SG.HUM PRF from school come
's/he has come back from school' (ibe203-00.270)

(3.97) (a) *ɔ́ á ʃiŋ̀*
3SG.HUM PROG walk
's/he is walking' (ibe027-00.222)

(b) *àu bá ʃiŋ̀ niŋ òjè mi iɲi èdɛ́n tu*
3SG.HUM.IDP FUT walk arrive place REL water that be
'he will walk to where the water is' (ibe301-00.022)

Table 3.40 displays the auxiliaries and their respective roles. The table also lists references to other examples in this sketch grammar where the respective auxiliary occurs.

Auxiliary	Position	Role	Example
das	postverb	goal	3.94 b
fàŋ	postverb	comparator	3.95 b
gbà	preverb	patient	3.194 b
gbà	preverb	instrument	3.89 a
kpa	postverb	beneficiary	3.158 b
niŋ	postverb	goal	3.96 a
ʃa	postverb	beneficiary	3.74 b
ʃa	postverb	recipient	3.104 b, c
ton	preverb	beneficiary	3.92 a
ton	preverb	partner	3.92 b
wa	preverb	source	3.96 b

Table 3.40: Thematic roles added by participant-adding auxiliaries

The auxiliary *ʃa*, which marks beneficiaries or recipients, is the base of the dative pronoun paradigm – probably a recent invention that is composed of *ʃa* and the respective object pronouns. The dative pronouns and their development is discussed in Section 3.5.2.

Adverbial auxiliaries These auxiliaries modify verbs and are thus called adverbial auxiliaries. Most of them have meanings traditionally grouped into the adverb word class in European languages. Others can be translated as English conjunctions or modal verbs. Depending on the kind of context they provide, they can be classified as temporal, directional, aspectual, modal, manner, causal, or comment auxiliaries. Some of them do not fall into any of these categories. The class of adverbial auxiliaries integrates many

Auxiliary	Position	Meaning	Category
ba	postverb	hither	directional
bɛ̀bɛ̀rɛ̀	preverb	fast	temporal
bɛ̀rɛ̀	preverb	start	aspectual
das	postverb	away	directional
deŋ	postverb	up	directional
dʒɔ	preverb	together	manner
gɛn	postverb	finish	aspectual
gbɛ́dɔ̀	preverb	must	modal
hɔ	postverb	away/out	directional
hɔ	postverb	too (much)	other
jànà	pre/postverb	early	temporal
jɔbɔ	preverb	can	modal
ki	preverb	still	temporal
kiǰ	postverb	back	directional
kɔ̃̀	preverb	just	other
mànà	preverb	almost	other
mĩ	preverb	?[11]	other
sisɛ̀n	preverb	just	temporal
sìsì	preverb	continue	aspectual
ʃáà	preverb	certainly	comment
ʃajo	preverb	first	temporal
ʃukù	postverb	together	directional
tɛtɛ	pre/postverb	slowly	temporal
tiɛ̀/tilɛ̀	preverb	actually	comment
tɔ̀n	preverb	be caused by	causal
wɛǹ	preverb	before	temporal

Table 3.41: Adverbial auxiliaries and their categorization

Yoruba loan words such as *dédé* 'accidently', or gbɛ́dɔ̀ 'must'. Table 3.41 lists all adverbial auxiliaries that appear more than once in the corpus.

Example 3.98 contains examples of three adverbial auxiliaries: *deŋ* 'up' (a), *ʃajo* 'first' (b), and tɛtɛ 'slowly' (c).

(3.98) (a) *ɔban èdɛ́n kpɛd deŋ*
child that jump climb
'that child jumped up' (ibe224-00.074)

(b) *òkú ìdín mi ʃajo ŋanà*
all those REL first germinate
'all of those that germinate first' (ibe167-00.095)

(c) *ɔ́ tɛtɛ á kɔ ìwé èdɛ́n*
3SG.HUM slowly PROG write letter that
'he wrote that letter slowly' (ibe178-01.169)

3.4.4.3 Defunct verbs

Abesabesi has three words that resemble verbs and take up a verbal position in a clause but feature a couple of deviant traits that distinguish them from prototypical verbs. As they can be categorized somewhere between verbs and particles, they will be called defunct verbs.

The first defunct verb is *gó*, which can roughly be translated as 'it is...'. Together with a preceding NP, it constitutes a separate utterance and forms the independent focus construction (see Section 3.8.1).

(3.99) *ònì gó*
1SG.IDP be
'it is me' (ibe346-00.140)

Gó never appears negated, as the negated equivalent is *ɛ́ɛ̀ tu ... è*, literally translated as 'it is not'. This construction uses the verbal copula *tu* and is discussed in Section 3.8.2.

The second defunct verb is *té* and can be translated as 'there is...'. It is preceded by an NP and forms an existential clause. Optionally, it can be followed by a dative object (participant-adding auxiliary *ʃa* plus object), an adjunct (NP or PP), or the negation particle *è*.

(3.100) (a) *àjè té sá*
permission exist 2SG.DAT
'you are allowed' (lit.: there is permission for you) (ibe104-00.140)

(b) *bábìʃ té èkiròm è*
<place_name> exist Ikaram NEG
'there is no Bar Beach in Ikaram' (ibe096-00.217)

[11]The verb *mĩ* has the meaning 'to do' as a full verb but also appears frequently as an auxiliary. Its function when used as an auxiliary could not be determined. If translated by consultants, clauses with the auxiliary *mĩ* and those without it do not differ. Very frequently, it appears in relative clauses introduced with (èdʒ)èdʒí mi (described in Section 3.7.4).

The third defunct verb is *sík͡ɔ́* 'where is/what about'. It is also preceded by an NP and may be used to inquire about a location (where is) or to lead a question towards another referent (what about). *sík͡ɔ́* does not appear with the negation particle *è*.

(3.101) (a) *iʃo* *so* *sík͡ɔ́?*
house 2SG.POSS where_is
'where is your house?' (ibe033-00.112)

(b) *ani* *sa* *sík͡ɔ́?*
people 2SG.POSS what_about
'what about your family?' (ibe017-00.079)

All of these three defunct verbs have a verbal form (CV or CVCV) but carry high tones. Lexical high tones are not frequent. High tones mostly are used as tonemes (see Section 3.1.5) or are carried by function words. Defunct verbs indeed rather have a grammatical function than a meaning. Furthermore, they never appear with TAM markers or bound subject pronouns. Instead, independent pronouns have to be used (as in Example 3.99). The preceding NP does not take a 1H (Section 3.1.5.2) as in affirmative verbal clauses.

3.4.5 Verb Phrase

A verb phrase (VP) minimally consists of its head – the verb. As discussed before, the verb can carry additional morphology such as bound subject pronouns, TAM prefixes, object pronouns, or the pluractionalis suffix. Depending on the verb's valency, it can be followed by no, one, or two obligatory objects. The object immediately following the verb is called the primary object and the other object is called the secondary object (Section 3.7.1). Object pronouns always represent the primary object, since they immediately follow the verb. In addition to these two objects, a dative object can occur. The dative object has developed from an auxiliary verb construction and is therefore situated at the end of a VP. A detailed description can be found in Section 3.4.5.1. The argument structure is discussed in Section 3.7.1.

The order of a verb phrase is thus:

$$PRO - TAM - V - PLUR + PO + SO + DatO$$

(3.102) (a) *PRO* *TAM* *V* *PO*
ɔ́ ka bɔ̀r atɔ
3SG.HUM PRF hoe ground
's/he hoed the ground' (ibe331-00.049)

(b) *PRO* *TAM* *V* *PO* *SO*
ǹ bá ki sa ɛmũ
1SG FUT give 2SG.OBJ money
'I will give you money' (ibe190-00.471)

(c)

PRO	*TAM*	*V*	*PO*	*DatO*
ɔ́	bá	gbà	iniŋ edʒe	ná
3SG.HUM	FUT	put	food	1SG.DAT

'she will give me food' (ibe039-00.046)

3.4.5.1 Auxiliary verbs

VPs with auxiliary verbs contain more than one verb and are thus usually more complex than mono-verbal VPs. As discussed in Section 3.4.4.2, auxiliary verbs can be preverbs or postverbs, depending on their position relative to the main verb. They can also be transitive or intransitive. The structure of a VP with auxiliary verb is thus:

$$PreV + PO + MainV + PO + SO + PostV + PO$$

(3.103) (a)

PRO	*PreV*	*MainV*	*PO*	*PostV*	*PART*
náà	gbɛ́dɔ̀	dʒe	iniŋ	hɔ	è
1SG.NEG	must	eat	thing	too_much	NEG

'I should not eat too much' (ibe111-00.133)

(b)

	MainV	*PO*		*PostV*	*PO*	*ADJUNCT*
oòsì,	mĩ	iniŋ	èŋèm	ʃa	ès	iʃo
God	do	thing	good	DAT	1PL.OBJ	house

'God, do us a favor in the house!' (ibe052-00.062)

As dative pronouns have developed from a postverbal auxiliary verb construction, they occupy a position after the main verb's objects. The construction *ʃa* 'DAT' + NP will be treated together with the dative pronouns as dative objects (DatO). Dative objects are obligatory in the ditransitive construction *gbà ʃa* 'give' and can thus be core arguments (see Section 3.7.1). In most cases, however, dative pronouns add a participant and are optional.

(3.104) (a)

PRO	*MainV*	*PO*	*DatO*
bà	gbà	ìwé	sá
3PL.HUM	put	book	2SG.DAT

'they gave you a book' (ibe089-00.010)

(b)

PRO	*MainV*	*PO*	*DatO*
bà	gbà	ìwé	ʃa u
3PL.HUM	put	book	3SG.HUM.DAT

'they gave her/him a book' (ibe089-00.011)

(c)

PRO	*MainV*	*PO*	*DatO*
bà	gbà	ìwé	ʃa ɔmis
3PL.HUM	put	book	DAT king

'they gave a book to the king' (ibe089-00.016)

3.4.5.2 Serial verb constructions

Serial verb constructions (SVCs), along with auxiliary verb constructions (Section 3.4.4.2) and sequential constructions (Section 3.4.2.2), are multi-verb clauses. An SVC is defined as a monoclausal construction where two main verbs are joined without any marker of coordination or subordination. The two main verbs of an SVC constitute two events that may occur simultaneously or consecutively. This distinguishes it from auxiliary-verb constructions (Sections 3.4.4.2 and 3.4.5.1), where only one main verb determines the event and one or multiple auxiliary verbs modify the event or add thematic roles to it. SVCs share the same temporal, aspectual, and modal setting unlike sequential constructions (Section 3.4.3).

(3.105) (a) *ísaj wo gbòm ɛ*
stone fall hit 3SG.NHUM.OBJ
'the stone fell and hit it' (ibe346-00.191)

(b) *náà gbɛ́dɔ̀ dʒe iniŋ hɔ è*
1SG.NEG must eat thing too_much NEG
'I should not eat too much' (ibe111-00.133)

(c) *ɔ́ ɲadum lí iʃum òdè, si á kàj ìwé*
3SG.HUM sit LOC on stool SEQU PROG read book
's/he sits on a chair and is reading a book' (ibe190-00.112)

Example 3.105 compares the three mentioned constructions. 3.105 a contains an SVC with two consecutive events, while 3.105 b shows the main verb *dʒe* modified by the two auxiliaries *gbɛ́dɔ̀* 'must' and *hɔ* 'too (much)'. The latter example merely contains one event – the eating. Clauses with adverbial auxiliaries, especially the directional postverbs, could theoretically be analyzed as SVCs. In other languages, constructions involving a motion verb to indicate the directionality of an event are usually analyzed as SVCs. However, here they are analyzed as auxiliaries, as they do not add a separate event but simply modify another. They therefore fit into the established category of auxiliary verbs. Example 3.105 c displays a sequential construction. In sequential constructions, the subject is shared. The two verbs each head their own VP and are part of two separate clauses. In contrast to SVCs, the consecutive clauses contain different TAM categories (in this example factative (unmarked) and progressive). Having distinguished SVCs from the similar auxiliary and sequential constructions above, the following paragraphs will now discuss the peculiarities of SVCs.

SVCs share at least one core argument. There are three possibilities of sharing a core argument in Abesabesi: sharing the subject, sharing the subject and an object, and the switch function where the object of the first verb is used as the subject of the second.

Many SVCs share only a subject. An example is 3.105 a from above. Others share the subject and the object. The shared object appears after the first transitive verb, while the following transitive verb has an empty object slot next to it (Example 3.106 a). Example 3.106 b shows a translation for the word identification card. The object of the first

verb *jas* ‘accept’ is the subject of the second verb *di* ‘enter’. This kind of switch function is also possible with ditransitive verbs. The primary object of the ditransitive verb *ki* ‘give’ in Example 3.106 c is the subject of the second verb *ʤe* ‘eat’, while the secondary object is the object of the second verb.

(3.106) (a) *ɔ́ ka kèd ɛ wos*
3SG.HUM PRF pick 3SG.NHUM carry_on_head
‘s/he has picked it up to carry it on the head’ (ibe332-00.452)

(b) *ìwé mi bà bàa jas ani di iʃo*
document REL 3PL.HUM use accept people enter house
‘a document they use to accept people entering the house’
(ibe111-00.092)

(c) *oɲo no á ki ɔban iniŋ edʒe ʤe*
wife 1SG.POSS PROG give child food eat
‘my wife is giving the child food to eat’ (ibe327-00.028)

Some SVCs seem to have lexicalized, as their meaning is not compositional or their verbal parts no longer have independent meaning. Examples are *di bò* ‘hold’ and *dò jè* ‘taste’. These cannot be analyzed as bisyllabic verbs or verb-noun compounds, as they have their primary object position in-between the two verbal parts.

3.5 Minor word classes

This section discusses all remaining word classes in their respective subsection: Determiners (Section 3.5.1), Pronouns (Section 3.5.2), Adverbs (Section 3.5.3), and Particles (Section 3.5.4).

3.5.1 Determiners

Determiners are used to refer back to referents already mentioned, to mark definiteness, or to indicate their distance from the speaker. Abesabesi determiners comprise four demonstratives and the determiner *ɛjɔ̀g* ‘another’.

The four demonstratives distinguish proximal from distal and singular from plural (see Table 3.42).

	SG	PL
proximal	èéni	èdɛ́n
distal	ìdín	ìdɛ́n

Table 3.42: Demonstratives

Demonstratives are used to distinguish objects close to the speaker from objects far from the speaker. They can be used as nominal modifiers after a noun or independently as

demonstrative pronouns (see Section 3.5.2). Example 3.107 shows the singular demonstratives used in one utterance where they are used to differentiate two soups. While the proximal demonstrative *èéni* is used as a modifier, the distal demonstrative *èdén* is used independently as a pronoun.

(3.107) *ìsɔ̀m èéni to fàŋ èdén*
soup this be_hot CPOR/SUP that
'this soup is hotter than that' (ibe147-00.133)

When used as modifiers, they agree with the modified noun in number (Section 3.3.2.1 and Example 3.108)

(3.108) (a) *ɔni èdén sì ʃu*
person that now run
'now that person ran' (ibe229-00.077)

(b) *ani ìdén bà sàg u kìj*
people those PST call 3SG.OBJ back
'those people called him back' (ibe120-00.022)

The determiner *ɛjɔ̀g* means 'another' and has the plural form *ijɔ̀g*. Like the demonstratives, it can be used independently as a pronoun or modify a noun.

Abesabesi has no articles and bare nouns can be definite or indefinite / specific or non-specific according to the context. Similar functions to articles might be taken by the numeral *ekìn* and the distal demonstratives, but this could not be investigated in detail. Example 3.109 shows two utterances of a speaker retelling the pear story he had been shown. The first protagonist, a man, is introduced with *owos ekìn* 'a man'. Later, the same man is referred to as *owos èdén* 'the man'. As these elements are not obligatory and bare nouns can be used in the same contexts, they will not be analyzed as a separate article word class.

(3.109) (a) *owos ekìn tu òdiɛ̀*
man one be there
'there was a man' (ibe041-00.002)

(b) *mi ɔ́ je ɛ kpé owos èdɛ́n je u è*
SIP 3SG see it COMP man that see 3SG.OBJ NEG
'he saw that the man did not see him' (ibe203-00.049)

Only very rarely, the low or mid toned *ɛ/ɛ̀* occurs in the corpus directly after nouns. As it only occurs in a few instances, it could not be investigated whether this is a shortened form of the demonstrative *èdén*, the 3SG.NHUM possessive pronoun *ɛ*, or an independent article.

(3.110) *ɛjau ɛ̀ tu iʃum iʃo*
monkey the be on house
'the monkey is on the house' (ibe015-00.082)

3.5.2 Pronouns

Pronouns are classified as a nominal category as they start with a vowel and can be the heads of NPs. This, however, only applies to free pronouns, as the three bound pronoun sets (subject, object, and possessive pronouns) attach to verbs or nouns. This section discusses all bound and free pronouns in Abesabesi. Table 3.43 lists all pronoun paradigms and refers to the respective section in which they are discussed.

Pronouns	Morpheme type	Section
Bound subject	Verbal prefix	3.5.2.1
Bound object	Verbal suffix	3.5.2.2
Bound possessive	Nominal suffix	3.5.2.3
Independent	Free	3.5.2.4
Logophoric	Free	3.5.2.5
Dative	Free	3.5.2.6
Independent Possessive	Free	3.5.2.7
Emphatic	Free	3.5.2.8
Indefinite	Free	3.5.2.9
Reciprocal	Free	3.5.2.10
Demonstrative	Free	3.5.2.11
Interrogative	Free	3.5.2.12

Table 3.43: Pronouns

Of these twelve pronoun sets, the first eight are personal pronouns (bound subject, bound object, bound possessive, independent, logophoric, dative, independent possessive, and emphatic). All personal pronouns, whether free or bound, distinguish three persons and two numbers (singular and plural). The third person in all personal pronoun types additionally distinguishes between human and non-human referents.

3.5.2.1 Bound subject pronouns

Bound subject pronouns are personal pronouns in subject position that always precede a verb. They only appear if there is no NP in subject position, which is why they cannot be analyzed as subject concordants. Bound subject pronouns can only be dropped in the affirmative imperative of the second person singular, with the sequential aspect, or when independent or emphatic pronouns are used. Although written as separate words (see orthographic conventions in Section 3.1.6), they are analyzed as verbal prefixes, since they adhere to the verb's VH. Abesabesi has three base sets of bound subject pronouns. For simplicity reasons the first set will be called realis pronouns and the second set will be called irrealis pronouns. As the realis/irrealis distinction is a controversial topic, it should be noted that this naming does not relate to a universal realis/irrealis category but has been chosen because the second set of pronouns is used with grammatical categories that

are sometimes associated with the irrealis label (negation, conditional). The third set are the habitual pronouns, which consist of the segmental part of the irrealis pronouns but a different melody. Combined with tonemes, and/or morphemes, these three pronoun sets make up seven paradigms. The QH, which seems to have its origin in a segmental morpheme, and the low toned nasal for the negative conditional paradigm, could in fact be analyzed as TAM morphemes, but as they differ from other TAM morphemes in form and position, they are treated here as part of the pronoun paradigms. Table 3.44 gives an overview of the seven paradigms, which will later be discussed in detail.

	declarative		interrogative	conditional
TAM	default	HAB	all	all
affirmative	realis + 1H	habitual +1H	QH + realis + 1H	irrealis + 1H
negative	irrealis		QH + irrealis	m̀ + irrealis

Table 3.44: Bound subject pronoun paradigms

The table shows that the pronoun paradigms are conditioned by the clause type (declarative, interrogative, and conditional), and by negation (affirmative and negative). Only in affirmative declarative clauses, the choice of paradigm is conditioned by the TAM category. The habitual marker *màá/mèé/mòó* triggers the habitual paradigm instead of the realis pronouns. All other paradigms are not triggered by TAM markers.

Realis pronouns The realis pronouns can be seen in Table 3.45. They are used for all affirmative clauses except for conditional clauses and for those with the habitual marker *màá/mèé/mòó*. All negated clauses and conditional clauses use the irrealis pronoun set.

The realis pronoun set has underlying tones. H for all pronouns of the third person except the 3PL.HUM and L for all other pronouns. The 3PL.HUM is the only pronoun of this set to harmonize with the verb. Realis pronouns are the base for the default affirmative paradigm and the affirmative interrogative paradigm (Table 3.46).

	SG	PL
1PERS	ǹ	à
2PERS	ɔ̀	àn
3PERS.HUM	ɔ́	bà/bè/bò
3PERS.NHUM	ɛ́	í

Table 3.45: Realis pronouns

Realis pronouns are most frequently used in the default paradigm, where they only carry the 1H. The 1H is pushed onto the following mora after an L. Otherwise, it merges with a preceding H (Section 3.1.5.2). As discussed in Section 3.1.6, the 1H is not written on the default affirmative subject pronouns.

(3.111) (a) *bà H sàg na* → *[bàságìna]*
3PL.SBJ 1H call 1SG.OBJ
'they called me' (ibe057-00.116)

(b) *ɔ́ H sàg na* → *[ɔ́sàgìna]*
3SG.SBJ 1H call 1SG.OBJ
's/he called me' (ibe001-02.035)

When used in the affirmative interrogative paradigm, they have an initial QH attached to the left end and the 1H attached to the right end. Instead of the QH, the morpheme *(mi)ní* can be used, which could be the source of the QH. However, it occurs very infrequently.

	SG	PL
1PERS	ń$^{\downarrow}$ń	á$^{\downarrow}$á
2PERS	ɔ́$^{\downarrow}$ɔ́	á$^{\downarrow}$ń
3PERS.HUM	ɔ́ɔ́	í$^{\downarrow}$bá/í$^{\downarrow}$bé/í$^{\downarrow}$bó
3PERS.NHUM	ɛ́ɛ́	íí

Table 3.46: Affirmative interrogative subject pronouns

On the realis pronouns with a high tone, this results in a lengthening of the vowel: QH + *ɔ́* + 1H → *ɔ́ɔ́* '3SG.HUM.Q'. On the low tone realis pronouns, this results in a downstep. The mono-moraic pronouns for the 1SG, 2SG, and 1PL are also lengthened with a downstep between the two resulting moras: for instance, QH + *ǹ* + 1H → *ń$^{\downarrow}$ń* '1SG.Q'. As the pronoun *àn* '2PL' has two moras, it is not lengthened, but the downstep occurs between the two moras: QH + *àn* + 1H → *á$^{\downarrow}$ń* '2PL.Q'. The pronoun *bà/bè/bò* '3PL.HUM' has an onset consonant and therefore places the QH on an epenthetic /i/: QH + *bà/bè/bò* + 1H → *í$^{\downarrow}$bá/í$^{\downarrow}$bé/í$^{\downarrow}$bó*. The initial epenthetic /i/ is a common strategy on nouns starting with a consonant (Section 3.1.3). As *ǹ* '1SG' is a syllabic nasal, the epenthetic /i/ does not have to apply to it. However, the form *í$^{\downarrow}$ń* also appears infrequently in place of *ń$^{\downarrow}$ń*.

(3.112) (a) *ɔ́$^{\downarrow}$ɔ́ ka ba?*
2SG.Q PRF come
'have you come?' (ibe005-00.192)

(b) *ɔ́ɔ́ ka ba?*
3SG.Q PRF come
'has s/he come?' (ibe005-00.264)

Irrealis pronouns Table 3.47 shows the irrealis pronouns, which are used for the affirmative habitual, the affirmative conditional, and all negative paradigms. They all exhibit a lenghtened vowel carrying a HL melody. The three-moraic pronouns *anáà* and *asáà* have an additional mid tone on the first mora. In fast speech, the lenghtened vowel carrying the HL melody is often reduced to a short vowel with a slightly falling mid

tone. Of the irrealis pronouns, all harmonize with the verb, except the 3SG.HUM, the 3SG.NHUM, and the 3PL.NHUM.

	SG	PL
1P	náà/néè/nóò	asáà/aséè/asóò
2P	sáà/séè/sóò	anáà/anéè/anóò
3P.HUM	ɔ́ɔ̀	báà/béè/bóò
3P.NHUM	ɛ́ɛ̀	íì

Table 3.47: Irrealis pronouns

The declarative negative paradigm, regardless of what TAM markers are present, uses the irrealis pronouns without any additional tones or morphemes (Example 3.113).

(3.113) (a) *asáà bá das ɛ̀ʃɔj è*
1PL.NEG FUT go tomorrow NEG
'we will not go tomorrow' (ibe190-00.083)

(b) *ɛ́ɛ̀ màá sabàá deŋ iʃum ísaj è*
3SG.NHUM.NEG HAB usually climb on stone NEG
'it did not use to climb on the rock' (ibe346-00.087)

The negative interrogative paradigm consists of irrealis pronouns and the QH. Instead of the QH, the morpheme *(mo)ní* is used – the negative equivalent of *(mi)ní* in the affirmative paradigm.

(3.114) (a) *H sáà kpas na è?*
QH 2SG.NEG hear 1SG.OBJ NEG
'don't you hear me?' (ibe186-00.145)

(b) *ní náà sɛmɛ̀ sa è?*
Q 1SG.NEG great 2SG.OBJ NEG
'did I not greet you?' (ibe265-00.267)

(c) *moní séè je íntiǹ è?*
Q.NEG 2SG.NEG see anything NEG
'have you not seen anything?' (ibe265-00.277)

The affirmative conditional paradigm is used in conditional or temporal clauses[12]. It consists of the irrealis pronouns and the 1H. The 1H docks onto the last vowel, lengthens it, and causes a downstep. For instance, *asáà* + 1H → *asáꜜá* '1PL.COND'.

(3.115) (a) *sáꜜá das iʃáafa*
2SG.COND go school
'if/when you go to school' (ibe168-00.002)

(b) *ɔ́ꜜɔ́ ŋɔ ɛ kpɛ́ré, ɔ́ bó hu*
3SG.HUM.COND drink it anyhow 3SG.HUM FUT die
'if she drinks it anyhow, she will die' (ibe241-00.056)

	SG	PL
1PERS	náˡá/néˡé/nóˡó	asáˡá/aséˡé/asóˡó
2PERS	sáˡá/séˡé/sóˡó	anáˡá/anéˡé/anóˡó
3PERS.HUM	ɔ́ˡɔ́	báˡá/béˡé/bóˡó
3PERS.NHUM	ɛ́ˡɛ́	íˡí

Table 3.48: Affirmative conditional subject pronouns

The negative conditional paradigm consists of a low toned syllabic nasal *m̀* and the irrealis pronouns. The syllabic nasal is pronounced [m] before vowels and [n] before the alveolar consonants /n/ and /s/. The 3PL.HUM seems to have an inherent /i/ in the beginning, which appears after the nasal: *mìbáà/mìbéè/mìbóò*.

(3.116) (a) *ǹséè dò mí ɔ̀ ʃénijes è*
2SG.NEG.COND want COMP 2SG be_fat NEG
'if you do not want to become fat' (ibe281-00.122)

(b) *m̀ɛ́ɛ̀ dàg è, ǹ bá gbà iniŋ edʒe ʃa ɛ*
3SG.NHUM.NEG.COND vomit NEG 1SG FUT put food DAT it
'if it does not vomit, I will feed it' (ibe346-00.031)

	SG	PL
1PERS	ǹnáà/ǹnéè/ǹnóò	m̀asáà/m̀aséè/m̀asóò
2PERS	ǹsáà/ǹséè/ǹsóò	m̀anáà/m̀anéè/m̀anóò
3PERS.HUM	m̀ɔ́ɔ̀	mìbáà/mìbéè/mìbóò
3PERS.NHUM	m̀ɛ́ɛ̀	m̀íì

Table 3.49: Irrealis pronouns

Habitual pronouns Declarative clauses where the habitual marker is used require a different set of subject pronouns: the habitual paradigm. It consists of the segmental part of the irrealis pronouns but with a different melody. Instead of the HL melody, the pronouns carry an M tone and the 1H. The complete paradigm can be seen in Table 3.50.

(3.117) (a) *neé mèé je u*
1SG.HAB HAB see 3SG.OBJ
'I used to see him' (ibe178-01.253)

(b) *baá màá ba ʃɔ́ʃì na*
3PL.HUM.HAB HAB come church 1SG.POSS
'they used to come to my church' (ibe108-00.093)

[12]As described in Section 3.7.4.3, conditional clauses can also have a temporal meaning.

	SG	PL
1	naá/neé/noó	asaá/aseé/asoó
2	saá/seé/soó	anaá/aneé/anoó
3.HUM	ɔɔ́	baá/beé/boó
3.NHUM	ɛɛ́	ií

Table 3.50: Affirmative habitual subject pronouns

3.5.2.2 Bound object pronouns

Object pronouns are verbal suffixes that have the following structures: 3SG pronouns and the 3PL.NHUM consist of only a vowel; 1SG and 2SG and 3PL.HUM have a CV structure, and the 1PL and 2PL suffixes have a short VC or the long VCVC form.

	SG	PL
1PERS	na/no	ès/àbès
2PERS	sa/so	èn/àbèn
3PERS.HUM	u	ba/bo
3PERS.NHUM	ɛ	i

Table 3.51: Object pronouns

The long forms of the 1PL and 2PL suffixes are identical with the respective independent pronouns (see Section 3.5.2). Both forms can be interchanged without change in meaning (Example 3.118).

(3.118) (a) *ɔ́ je àbès*
3SG.HUM see 2PL.OBJ
's/he sees us' (ibe121-00.239)

(b) *ɔ́ je ès*
3SG.HUM see 2PL.OBJ
's/he sees us' (ibe005-00.250)

Only the suffixes with a CV structure adhere to VH: *na/no* '1SG.OBJ', *sa/so* '2SG.OBJ', and *ba/bo* '3PL.OBJ'. For a detailed discussion on VH and examples, see Section 3.1.4.4 and Table 3.19.

The object pronoun paradigm is identical to the possessive pronoun paradigm (see Section 3.5.2.3) and likewise triggers the GenH.

3.5.2.3 Bound possessive pronouns

The possessive pronoun paradigm consists of nominal suffixes and is identical in form to the object pronoun paradigm (see Section 3.5.2.2). Possessive pronouns trigger the

GenH on the preceding noun. The 3SG.HUM, 3SG.NHUM and 3PL.NHUM pronouns consist of only a vowel. The 1SG, 2SG, and 3PL.HUM pronouns have a CV structure and are the only forms to adhere to VH. The 1PL and 2PL pronouns have a long VCVC and a short VC structure. Long and short forms are interchangeable without change in meaning (Example 3.119). The long forms are identical to the respective independent pronouns.

	SG	PL
1PERS	na/no	ès/àbès
2PERS	sa/so	èn/àbèn
3PERS.HUM	u	ba/bo
3PERS.NHUM	ɛ	i

Table 3.52: Possessive pronouns

(3.119) (a) *ʃum àbèn*
head 2PL.POSS
'your heads' (ibe121-00.171)

(b) *ʃum èn*
head 2PL.POSS
'your heads' (ibe005-00.180)

3.5.2.4 Independent pronouns

This paradigm contains pronouns that are prosodically, morphologically, and syntactically independent. The pronouns all start with the vowel /ò/ or /à/ followed by a part that is similar in form to the bound possessive/object pronouns (Sections 3.5.2.3 and 3.5.2.2). While the first and second person pronouns carry only low tones, the third person human pronouns carry a low and a mid tone. Independent pronouns never adhere to vowel harmony.

	SG	PL
1PERS	ònì	àbès
2PERS	òsì	àbèn
3PERS.HUM	àu	àba
3PERS.NHUM	ɛ̀jì	ìjì

Table 3.53: Independent pronouns

These pronouns are generally used to express emphasis. A pronoun in the fronted focus position, for example, must be an independent pronoun. In this position, the pronoun can function as subject, primary object, or auxiliary object (see Example 3.120). Example 3.120 c shows a fronted pronoun that is the object of the auxiliary verb *ʃa* (auxiliary verbs are discussed in Section 3.4.4.2).

(3.120) (a) *ònì gí ʃag atɔ*
1SG.IDP FOC sweep floor
'it was me who swept the floor' (ibe346-00.160)

(b) *ònì gí bà kud dʒo*
1SG.IDP FOC 3PL first give_birth
'it is me they gave birth to first' (ibe065-00.066)

(c) *ònì gí ɔ́ mĩ̀ ɛ̀ɲèm ʃa*
1SG.IDP FOC 3SG do favor DAT
'it is me he did a favor for' (ibe365-00.081)

Independent pronouns can also be used as subject pronouns without the focus marker *gí* and replace the bound subject pronoun (see Example 3.121). In this case, without the focus marker *gí*, they can only function as subjects and must carry the 1H (see Section 3.1.5.2). A semantic difference between Example 3.121 and 3.120 a is not perceived by speakers. In both cases, the pronoun is emphasized.

(3.121) *ònì H ʃag atɔ → [ònɪ́ ʃagatɔ]*
1SG.IDP 1H sweep floor
'I swept the floor' (ibe346-00.158)

Besides the fronted focus position, there are other contexts where independent pronouns must be used instead of bound pronouns. Firstly, they are required with the conjunction *íni* or its Yoruba equivalent *àti*. Only the emphatic pronoun can also occur in that position (see Example 3.122 c).

(3.122) (a) *ònì íni ohij no […]*
1SG.IDP and older_sibling 1SG.POSS
'me and my older sibling' (ibe203-00.178)

(b) *[…] àu àti owos u*
3SG.HUM.IDP and husband 3SG.POSS
'her and her husband' (ibe065-00.182)

(c) *bò bug àti ani u àti àtɛ̀n u*
3PL.HUM beat and people 3SG.POSS and EMPH 3SG.POSS
'they beat his family and himself' (ibe301-00.135)

Secondly, they have to be used after certain function nouns, such as *èdʒì* 'like', *ègido* 'only', and *íʃa* 'for' (Example 3.123).

(3.123) (a) *ɔni èdʒì àu té è*
person like 3SG.HUM.IDP exist NEG
'there is no person like him/her' (ibe140-00.050)

(b) *ègido ònì*
only 1SG.IDP
'only me' (ibe265-00.053)

(c) *íʃa* *àbès* *àgbɛ̀*
for 1pl.IDP farmer
'for us farmers' (ibe167-00.005)

Thirdly, they have to be used before defunct verbs (see Section 3.4.4.3).

(3.124) *ònì* *gó*
1SG.IDP be
'it is me' (ibe346-00.140)

And lastly, only independent pronouns can be modified by relative clauses or numerals.

(3.125) (a) *iʃo* *àba* *mi* *bé* *moŋo* *ijel* *èéni*
house 3PL.HUM.IDP REL FUT marry year this
'the house of them who want to marry this year' (ibe241-00.118)

(b) *àba* *ídiàn*
3PL.HUM.IDP two
'two of them' (ibe140-00.258)

The forms *àbès* '1PL.IDP' and *àbèn* '2PL.IDP' are identical with the long forms of the 1PL and 2PL bound possessive and object pronouns (Sections 3.5.2.3 and 3.5.2.2). The independent pronouns *àu* '3SG.HUM.IDP' and *èjì* '3SG.NHUM.IDP' are also used as logophoric pronouns (see Section 3.5.2.5).

3.5.2.5 Logophoric pronouns

Abesabesi uses the 3SG independent pronouns as logophoric pronouns: *àu* '3SG.HUM.LOG' and *èjì* '3SG.NHUM.LOG'. Logophoric pronouns are a means of facilitating reference tracking in complex sentences. Abesabesi uses logophoric pronouns in all complement clauses where the third person of the matrix clause has the same referent as the embedded clause. For instance, in Example 3.126 a, a person says about him/herself, that s/he is writing a letter. The complement clause's 3SG subject pronoun must be replaced by the third person logophoric pronoun *àu* because the referent of the matrix clause's and complement clause's subject are the same. In Example 3.126 b, on the other hand, a person says that another person is writing a letter. The referents are not the same and thus, the usual 3SG subject pronoun is used (*ɔ́*).

(3.126) (a) *ɔ́* *mò* *àu* *á* *kɔ* *lɛ́tà*
3SG.HUM say 3SGHUM.LOG PROG write letter
'S/he$_i$ said s/he$_i$ is writing a letter' (ibe203-00.090)

(b) *ɔ́* *mò* *mí* *ɔ́* *á* *kɔ* *lɛ́tà*
3SG.HUM say COMP 3SGHUM PROG write letter
'S/he$_i$ said s/he$_j$ is writing a letter' (ibe203-00.096)

Logophoric pronouns are used when the matrix verb is a verb of speech, thought, knowledge, direct perception or desire. The logophorics are also used in subsequent clauses if the indirect speech continues. Examples 3.127 a-c show three consecutive utterances of a story. The utterance in Example 3.127 c is continued indirect speech, where the logophoric pronoun *àu* is used. The example also shows that object pronouns do not have a logophoric form. In 3.127 a, the object of *ton* 'meet' is the regular object pronoun, although it has the same referent as the subject of the matrix clause.

(3.127) (a) *[...] ɔ́ mò mí bà ton u á wuʃ*
3SG.HUM say COMP 3PLHUM meet SG.OBJ PROG collect
iŋni èdɛ́n dìg hɔ
water that pour away
'he said they should help him to remove the water' (ibe301-00.020)

(b) *bá'á wuʃ iŋni dìg hɔ*
3PL.HUM.COND collect water pour away
'that if they remove the water' (ibe301-00.021)

(c) *àu bá ʃiŋ niŋ òjè mi iŋni èdɛ́n tu*
3SG.HUM.LOG FUT walk GOAL place REL water that be
'he would walk into where the water was' (ibe301-00.022-023)

There is also a logophoric pronoun for the 3SG.NHUM – *èji* (Example 3.128). As animals or inanimate entities rarely speak, know, perceive, or want something, it does not occur very often in the corpus. Plural logophors do not exist (see Example 3.129).

(3.128) *ɛ́ dò mí èji dàg*
3SG.NHUM want COMP 3SG.NHUM.LOG vomit
'it wants to vomit' (ibe346-00.022)

(3.129) *bè dò mí bà miʃ*
3PL.NHUM want COMP 3PL.NHUM sleep
'they want to vomit' (ibe346-00.220)

3.5.2.6 Dative pronouns

The dative pronouns are a paradigm of personal pronouns that take the thematic roles of a beneficiary or recipient. This paradigm appears to be a fairly new development, as only two forms have a contracted form and the rest consists of the participant-adding auxiliary *ʃa* (see Section 3.4.4.2), the GenH, and the object pronouns. Table 3.54 shows all forms of the paradigm with pronunciation in square brackets.

All forms except the 1SG and 2SG cannot really be seen as pronouns but as combinations of an auxiliary, the GenH[13], and object pronouns. Their behavior does not differ from other auxiliary constructions (Example 3.130).

[13]As the functional tones are not transcribed (see Section 3.1.6), the GenH does not occur in the phonemic forms of Table 3.54, but appears in the phonetic forms.

	SG	PL
1PERS	ná	ʃa ès [ʃes]
2PERS	sá	ʃa èn [ʃeǹ]
3PERS.HUM	ʃa u [ʃáu]	ʃa ba [ʃába]
3PERS.NHUM	ʃa ɛ [ʃéɛ]	ʃa i [ʃíi]

Table 3.54: Dative pronouns

(3.130) *bà ta iʃo ʃa u*
3PL.HUM build house DAT 3SG.OBJ
'they build her/him a house' (ibe073-00.183)

The reason these forms are analyzed as a pronoun paradigm are the forms of the 1SG and 2SG. By following the regular pattern *ʃa* + GenH + bound object pronoun, we would expect the 1SG and 2SG forms to be *ʃa na* [*ʃána*] and *ʃa sa* [*ʃásа*]. These, however, are judged as ungrammatical by native speakers (Example 3.131 a). Instead, the forms *ná* and *sá* are used (Example 3.131 b). These forms are likely to have developed from deleting the auxiliary *ʃa*, leaving only the object pronouns *na* and *sa*, but with the GenH replacing the mid tone on the object pronouns. The high tone is in fact the only component that distinguishes these two dative pronouns from bound object pronouns. The invariable /a/ is indicative of former VH with the deleted auxiliary *ʃa* – the resulting pronouns do not harmonize. As the dative pronouns originated in a verbal construction, they are the only independent pronouns to have no initial vowel.

(3.131) (a) * *ɔ̀ gbà ìwé ʃa H na*
2SG put book DAT GenH 1SG.OBJ
'you gave me the book' (ibe033-00.045)

(b) *ɔ̀ gbà ìwé ná*
2SG put book 1SG.DAT
'you gave me the book' (ibe033-00.044)

Their origins as a combination of *ʃa* and the respective object pronoun can be seen when they are fronted into the focus position. Example 3.132 a shows a sentence with the 1SG.DAT pronoun *ná*. When focus-fronted, it appears as an independent pronoun in the front, and at its original spot, the auxiliary *ʃa* remains (Example 3.132 b).

(3.132) (a) *ɔ́ mĩ̀ èŋèm ná*
3SG do favor 1SG.DAT
's/he did me a favor' (ibe365-00.071)

(b) *ònì gí ɔ́ mĩ̀ èŋèm ʃa*
1SG.IDP FOC 3SG do favor DAT
'it is me s/he did the favor for' (ibe365-00.081)

3.5.2.7 Independent possessive pronouns

Besides the bound possessive pronouns (see Section 3.5.2.3), Abesabesi has independent possessive pronouns. These can be used without a noun that expresses the possessed (Example 3.133).

(3.133) *iʃo* *ebòn* *tu* *èsó*
house which be yours
'which house is yours?' (ibe033-00.099)

Table 3.55 presents the paradigm. The 3PL.NHUM pronoun could not be elicited and did not appear in the corpus. The other forms seem to have simple and reduplicated forms. The reduplicated form appeared more frequently. Only some pronouns' simple forms appeared in the corpus. The others might not exist or are lacking in the corpus.

	SG	PL
1PERS	ènó / ènóònó	èŋeŋès
2PERS	èsó / èsóòsó	èŋeŋèn
3PERS.HUM	èjóòjóu	èjóbo / èjéèjobo / èjóòjóbo
3PERS.NHUM	èjéèjɛ́ɛ	?

Table 3.55: Independent possessive pronouns

3.5.2.8 Emphatic pronouns

This pronoun paradigm is composed of the morpheme *àtɛ̀n* and the respective possessive pronouns. The 1PL and 2PL pronouns seem to have shortened forms (*àtàbes* and *àtàbèn*).

	SG	PL
1PERS	àtɛ̀n na [àtɛ̀nna]	àtɛ̀n àbès [àtɛ̀nàbès]/ àtàbès [àtàbès]
2PERS	àtɛ̀n sa [àtɛ̀nsa]	àtàbèn [àtàbèn]
3PERS.HUM	àtɛ̀n u [àtɛ̀nu]	àtɛ̀n ba [àtɛ̀mba]
3PERS.NHUM	àtɛ̀n ɛ [àtɛ̀nɛ]	?

Table 3.56: Emphatic pronouns

The emphatic pronouns can be used as subject pronouns or after conjunctions such as *àti* 'and' (Example 3.134 a, b). They are used for emphasis and translate as 'myself, yourself, etc.'. Very often, they occur in subject position modified by the Yoruba loan nɔ́ɔ̀, which means 'also' (Example 3.134 c). *Nɔ́ɔ̀* cannot modify bound subject pronouns. It only occurs with emphatic pronouns.

(3.134) (a) *àtɛ̀n* *u* *sisɛ̀n* *bò* *kùʃo*
EMPH 3SG.POSS just PST descend
'he himself had just come down' (ibe041-00.067)

(b) *dibén bò bug àti ani u àti àtèn u*
3PL.HUM beat and people 3SG.POSS and EMPH 3SG.POSS
'now, they beat his family and himself' (ibe301-00.135)

(c) *àtèn u nɔ̋ɔ̀ sɛmɛ̀ u è*
EMPH 3SG.POSS also greet 3SG.OBJ NEG
's/he also did not greet him/her' (ibe265-00.350)

àtèn might be a noun, but it never occurs without modifiers in the corpus. It only occurs with possessive pronouns or nouns. Nouns can also be emphasized by *àtèn*. Assuming *àtèn* is a noun, it functions as the head of an NP and the emphasized noun which follows then acts as a modifier (Example 3.135).

(3.135) *àtèn idʒàkpá bà das*
EMPH tortoise PST go
'the tortoise himself left' (ibe301-00.079)

3.5.2.9 Indefinite pronouns

The generic nouns *ɔni* 'person/somebody', *iniŋ* 'thing/something' and *òjè* 'place/somewhere', sometimes in combination with the numeral *ekìn* 'one', are also used as specific indefinite pronouns

(3.136) (a) *iniŋ mi ó bùr ɔni ba èhen*
something REL PROG chase somebody come now
'something that runs after somebody' (ibe140-00.113)

(b) *iji ekìn, ǹ tu òjè ekìn*
day one, 1SG be place one
'One day, I was somewhere' (ibe140-00.207)

Their non-specific indefinite counterparts are formed by partial reduplication and the insertion of the morpheme *-ńt-* (Section 3.6.4): *ɔńtɔn* 'anybody/nobody/whoever', *ińtiǹ* 'anything/nothing/whatever', and *òńtòjè* 'anywhere/nowhere/wherever'.

(3.137) (a) *íntiǹ mi ɔ́ ʃa, ɔ́ ba ʃɛǹ è*
anything REL 3SG.HUM say, 3SG.HUM come speak NEG
'whatever she says, he does not answer' (ibe168-00.149)

(b) *òntójè mi ɔ́ das, ɔ́ɔ̀ bá ʃi u*
anywhere REL 3SG.HUM go 3SG.HUM.NEG FUT leave 3SG.OBJ
hɔ è
away NEG
'wherever she goes, he will never leave her' (ibe140-00.207)

These non-specific indefinite pronouns have a negative indefinite reading when occurring in a negated clause.

(3.138) (a) *íntiǹ tu ìwùd ìtog èdén è*
nothing be inside calabash that NEG
'there is nothing in that calabash' (ibe147-00.075)

(b) *néè je ɔ́ntɔn è*
1SG.NEG see anybody NEG
'I did not see anybody' (ibe65-00.039)

(c) *néè je ɛ ɔ́ǹtòjè è*
1SG.NEG see 3SG.NHUM.OBJ anywhere NEG
'I could not find it anywhere' (ibe65-00.096)

The numeral *ekìn* can also be reduplicated in the same way to form *eńtekìn* 'anything/nothing/whichever'. It has a similar meaning to *iǹtiǹ*, but can also be used as a modifier for other nouns.

(3.139) (a) *bé^{l}é mĩ eńtekìn, ɔ́ bá sà*
3PL.HUM.COND do anything, 3SG.HUM FUT know
'if they do anything, s/he will know' (ibe140-00.140)

(b) *néè kinè je ebo eńtekìn èkiròm òhun è*
1SG.NEG PRF.NEG see dog any Ikaram here NEG
'I have not seen any dogs in Ikaram' (ibe292-00.575)

3.5.2.10 Reciprocal pronoun

Abesabesi's word for 'body' *ijes* can be used as a reciprocal pronoun (Example 3.140 a). Sometimes, the 3PL.HUM possessive pronoun is added (Example 3.140 b). If used with other possessive pronouns, it can function as a reflexive pronoun (Example 3.140 c). Other meanings as a functional noun were discussed in Section 3.3.3.3.

(3.140) (a) *bà sɛmɛ̀ ijes*
3PL.HUM greet RECP
'they greeted each other' (ibe057-00.082)

(b) *sí á sɛmɛ̀ ijes bo*
SEQU PROG greet RECP 3PL.POSS
'they are greeting each other' (ibe008-00.028)

(c) *ǹ bá wol ijes no*
1SG FUT kill REFL 1SG.POSS
'I will kill myself' (ibe057-00.089)

3.5.2.11 Demonstrative pronouns

The demonstratives introduced in Section 3.5.1 can also be used as demonstrative pronouns. In this function, they do not modify a noun, but are used independently and constitute a noun phrase.

(3.141) (a) *èéni gí à sàg ịni*
this FOC 1PL call water
'we call this water' (ibe017-00.208)

(b) *màd tu ìdén*
what be those
'what are those?' (ibe365-00.100)

3.5.2.12 Interrogative pronouns

The interrogative pronouns can be seen in Table 3.57. A common structure is only detectable on the pronouns *màd*, *máàn*, and *mèj*. They all start with an /m/ and consist of one closed syllable. *màd* and *mèj* each have a variant starting with a /b/ instead of an /m/. *ímèj* is merely a combination of *mèj* and the LocH. As *mèj* has no initial vowel, an epenthetic /i/ is added where the LocH can dock (see Example 3.34 c). Thus, when the pronoun *mèj* is used to ask for a location instead of a direction, it must appear with the LocH.

Pronoun	Meaning
tí	what
màd/bàd	what/which
ebòn	which
máàn	who/whom/whose
kíjé	where
ímèj/íbèj	where (location)
mèj/bèj	where (direction)
ìnɛ́	how much/how many

Table 3.57: Interrogative pronouns

All these pronouns must be part of the sentence's first phrase and therefore carry the 1H if used in a sentence (Example 3.142 a). *màd* 'which', *máàn* 'whose', *ebòn* 'which', *mèj* 'from where', and *ìnɛ́* 'how much/how many' can be used to modify a noun which they follow (Example 3.142 b).

(3.142) (a) *màd H tu èdén → [màdí tèdén]*
what 1H be that
'what is that?' (ibe017-00.201)

(b) *ìbùrɛ ebòn H bà ba → [bùrebòníbàba]*
time which 3PL.HUM come
'when did they come?' (ibe121-00.251)

3.5.3 Adverbs

This word class is similar to nominals, as all adverbs start with a vowel. Adverbs might have developed from nouns, which would explain the initial vowel, but as they cannot

head an NP or function as a core argument of a verb, they cannot be counted as nominals. Adverbs are categorized by a low tone on the initial vowel (with the exception *ijɛ̀n* 'yesterday'). They all give a deictic location, time, or manner reference. The following list contains all adverbs found in the corpus.

- Adverbs of time
  - *èhen* 'now'
  - *ìdʒɛ* 'today'
  - *ijɛ̀n* 'yesterday'
  - *ɛ̀ʃɔj* 'tomorrow'
  - *ìdomog* 'this year'
  - *àmaj* 'next year'
  - *èkìdɛ* 'in the old times'
  - *òʃé* 'in the old times'
- Adverbs of place
  - *ìhun/òhun* 'here'
  - *òden* 'here'
  - *òdiɛ̀* 'there'
  - *ìjen* 'there'
  - *èsè* 'outside'
- Adverbs of manner
  - *èren* 'so'

Adverbs cannot be modified. They usually take the default spot for adjuncts after the VP (Example 3.143 a). Moreover, they can appear in the fronted focus position (Example 3.143 b).

(3.143) (a) *asáà* *bá* *das* *ɛ̀ʃɔj* *è*
1PL.NEG FUT go tomorrow NEG
'we will not go tomorrow' (ibe190-00.083)

(b) *ìdʒɛ* *gí* *tu* *èʃî* *èkiròm*
today FOC be market Ikaram
'today is market day of Ikaram' (ibe140-00.036)

Adverbs of place must appear with LocH if they are used to mark a location. If they are used to mark a direction, the LocH does not appear.

3.5.4 Particles

This final word class contains miscellaneous words that can neither be affiliated to the group of nominals nor to the group of verbs. They all are short words with a grammatical or pragmatic function.

Particle	Meaning	Reference
lí/ʃí	LOC	Section 3.2.2.3
mi	REL	Section 3.7.4.2
mí	COMP	Section 3.7.4.1
mi	SIP	/
(mi)ní	Q	Section 3.7.5.1
(mo)ní	Q.NEG	Section 3.7.5.1
gí	FOC	Section 3.8.1
ŋ̀gí	CFOC	Section 3.8.4
è	NEG	Section 3.7.6
áj̀	Q	Section 3.7.5.1
ò	EMPH	/
íni/àti	and	/

Table 3.58: Particles

Table 3.58 contains an exhaustive list of particles (apart from interjections). It shows that the particle word class contains words with diverse functions: the prepositions *lí* and *ʃí*, conjunctions such as the relativizer *mi* and the complementizer *mí*, the focus particle *gí*, the negation particle *è*, the question particle *áj̀*, the emphatic particle *ò*, the correlative conjunctions *íni* and *àti*, and the sentence-initial particle (SIP) *mi*, whose function is not yet clear. Each particle has a specific position in the sentence. While the SIP introduces a sentence and the relativizer *mi* and the complementizer *mí* introduce relative and complement clauses respectively, the sentence-final particles *è*, *áj̀*, and *ò* only occur at the end of a sentence. Apart from the particles in Table 3.58, the word class of particles also contains various interjections such as *ŋ́kɔ̀* 'thanks/please', *hɛ̃ɛ̃/ɛ̃ɛ̃/hɛ̃́ɛ̃́* 'yes', and *héèjè/éèjè* 'no'.

3.6 Word formation

This section deals with morphological devices to create new words. Abesabesi employs three main strategies: derivation, composition, and reduplication. Derivation only exists in the form of deverbal nominalization, which is discussed in the following subsection.

3.6.1 Deverbal nominalization

Nominalization of verbs is used either to create new concepts or to make verbs accessible to NP positions in the clause. Abesabesi has two strategies to nominalize verbs: Simple nominalization and complex nominalization.

3.6.1.1 Simple nominalization

This strategy is the simpler one, as only a vowel is prefixed in order to match the form of a noun. The quality of the vowel is not determinable. Many of the words created by simple nominalization may have been coined when the noun class system was still productive. The quality of the vowel therefore indicated the affiliation to a noun class. These words include, for example, *ɔʃu* 'a run' from *ʃu* 'to run', *ɛkɔn* 'a fight' or *ikɔn* 'a war' from *kɔn* 'to fight', and *ohu* 'death' from *hu* 'to die'. Other words created by this strategy, however, are likely to be newly coined words:

(3.144) (a) *ɔ̀-faǹ-àg*
NMLZ-knead-PLUR
'bread' (ibe190-00.124)

(b) *ò-gbegel*
NMLZ-roll
'car' (ibe075-00.020)

(c) *o-tì-lí-atɔ* → *[otìlátɔ]*
NMLZ-push-LOC-ground
'pen' (ibe039-00.082)

(d) *ɔ̀-hùr-H-òdùg* → *[ɔ̀hùródùg]*
NMLZ-put_inside-LocH-leg
'shoe' (ibe298-00.377)

The examples in Example 3.144 show that verbal extensions (the pluractionalis extension in Example 3.144 a), or adjuncts (Example 3.144 c, d) can be incorporated. These recently coined words all have /o/ or /ɔ/ as the initial vowel, depending on the ATR value of the verb. Yet, there are not many of these words, as new concepts are usually borrowed from Yoruba or English. Speakers stated that Abesabesi equivalents are only rarely coined. If they are coined, they are used so that people who don't understand Abesabesi will not be able to understand the topic of the conversation.

The results of this nominalization process are thus abstract verbal nouns or concrete objects. They are not different from any other regular noun and can be used in the same syntactic context or be modified.

3.6.1.2 Complex nominalization

This strategy involves the discontinuous morpheme *i- -oǹ/i- -aǹ* and results in a form that will henceforth be called the "gerund". The circumfix consists of a prefixed /i/ and

the suffix *-oǹ/-aǹ*, which harmonizes according to the suffix-VH (Section 3.1.4.4). The prefixed /i/ seems to be optional, as it is sometimes dropped.

Verb	Meaning	Gerund
de	buy	idòoǹ
ku	fall	ikòoǹ
là	plant	ilàaǹ
lɔ	throw	ilàaǹ
bèlè	take care	ibèloǹ
wuʃò	plant	iwùʃoǹ
bɔgɔ̀j	shout	ibɔ̀gɔ̀jaǹ
ɲan	step on	iɲànaǹ

Table 3.59: Gerund forms

Table 3.59 shows that the verbal stem, regardless of the lexical tone, carries low tones in the gerund form. The gerund thus always has a MLML melody. The final vowel in CV verbs is not deleted in the gerund form in order to maintain the MLML melody. Yet, it is completely assimilated to the /o/ or /a/ of the circumfix. This and the deletion of the lexical tones result in many homonymous forms. *idòoǹ* for example is the gerund for the verbs *dò* 'want' and *de* 'buy'. Final vowels in polysyllabic verbs are deleted before the suffixed part of the gerund: *bèlè* 'take care' becomes *ibèloǹ*.

The term "gerund" has been adopted from (Elugbe, 1984, p. 88), who uses it for a similar nominalization form in the Edoid languages. He reconstructs the circumfix as *U-... -AmhI in Proto-Edoid, which is probably a cognate form of the Abesabesi gerund.

The gerund is lexicalized and used as a noun or part of a noun in two instances in the natural corpus: *imùnaǹ* 'new yam festival' from *munà* 'celebrate the new yam' or *ìguj̀ iwùʃoǹ* 'seed tubers' from *ìguj̀* 'materials' and *wuʃò* 'to plant'. The other three times, it occurs in a construction that is henceforth called the "relativized focus construction".

(3.145) (a) *ikòoǹ mi ɔ́ ku èdén*
fall\GRND REL 3SG.HUM fall that
'the falling he fell' (ibe229-00.053)

(b) *idàsaǹ mi ɔ́ das èdén*
go\GRND REL 3SG.HUM go that
'the going he went' (ibe229-00.051)

(c) *nàbùnaǹ mi ɔ́ nàbù ni èdén*
stand_up\GRND REL 3SG.HUM stand_up ? that
'the standing up he stood up' (ibe065-00.041)

All three examples have the same structure:

$$Gerund + Relativizer + Subject + ResumptiveVerb + Demonstrative$$

The gerund is followed by a relative clause and a demonstrative. The relative clause contains only a subject and the same verb as in the gerund. The translations given in the above example are literal ones given by the consultant. The function seems to be an emphasis of the event. This can be seen in the context of the first example (Example 3.146). It describes how a boy fell. After a declarative sentence that introduces the event of the boy falling (a), the relativized focus construction occurs (b). The following sentence (c) then states the result of the falling.

(3.146) (a) *mi ɛ́ be niŋ ìbùrɛ ekìn mi ɔ́ ku*
SIP 3SG.NHUM PST reach time one REL 3SG.HUM fall
'suddenly, he fell' (ibe229-00.052)

(b) *ikòoǹ mi ɔ́ ku èdɛ́n*
fall\GRND REL 3SG.HUM fall that
'the falling he fell' (ibe229-00.053)

(c) *mi ɔ́ ʃiga gbà atɔ*
SIP 3SG.HUM rise hit ground
'he hit the ground' (ibe229-00.054)

In an elicitation session, a consultant confirmed that the gerund can be used productively in this construction. In two follow-up elicitations (ibe361 and ibe362), he produced this construction with a set of over 200 verbs. According to him, the gerund could also be replaced by the abstract verbal nouns (simple nominalization) (Example 3.147 a) or by a form that is probably the Yoruba nominalization through reduplication (Example 3.147 b).

(3.147) (a) *ɔ-ʃu mi ɔ́ á ʃu èdɛ́n*
NMLZ-run REL 3SG.HUM PROG run that
'that run s/he ran' (ibe349-00.277)

(b) *bí-ba mi ɔ́ ba èdɛ́n*
NMLZ-come REL 3SG.HUM PROG come that
'that coming s/he came' (ibe349-00.264)

The Yoruba nominalization through reduplication is produced by reduplicating the initial consonant of a verb before a high toned /í/: *rà* 'to buy' – *rírà* 'buying'. (Sachnine, 2014, p. 62) This nominalization technique is used in Yoruba when the verb is focus fronted.

(3.148) *rírà ni mo ra ìwé*
buying be I buy book
'It's buying that I bought the book' Yoruba (Yusuf, 1990, p. 89)

As Example 3.148 shows, the nominalization technique and its use are very similar. Therefore, it is assumed that the nominalization strategy used in Example 3.147 b is borrowed from Yoruba.

Agoyi (2008, p. 72) mentions the gerund and that it is found "when verbs are preposed in focus construction in Àbèsàbèsì". It is not clear if she refers to the relativized gerund construction.

3.6.2 Composition

Composition still lacks a structured analysis and is not yet fully understood. There are several constructions that could be analyzed as compound nouns. Firstly, there are noun-noun combinations that behave exactly like possessive constructions. This means a noun is connected through the GenH with the possessor noun. Only their figurative meaning and their frequent occurrence in this combination distinguishes them from possessive constructions. Examples are *ɛmű́ʃum* 'bride price' from *ɛmũ* 'money' and *iʃum* 'head', or body parts such as *ɔhɔ̃́bɔ* 'elbow pit' from *ɔhɔ̃* 'neck' and *ɔbɔ* 'arm'. A second construction are noun-noun connections with a GenH but where the first vowel is assimilated instead of deleted. This is usually only detectable in careful speech. In fast speech, this construction resembles the possessive construction. Examples are *iʃáafa* 'school' from *iʃo* - 'house' and *afa* 'leaf/book', and *onóokpo* 'front yard' from *onu* 'mouth' and *okpo* 'way'. The third construction is a noun-noun connection without a GenH. Examples are various compounds with *ɔni* 'person', such as *ɔniʃum* 'chairman/boss' from *ɔni* 'person' and *iʃum* 'head', or *ɔnafa* 'herbalist' from *ɔni* 'person' and *afa* 'leaf'.

As the GenH only appears on M tone moras, it is impossible to associate a noun-noun connection with low tones to one of these three constructions.

3.6.3 Object incorporation

Object incorporation as a word formation process is impossible to distinguish from the phonemic process of vowel deletion (Section 3.1.4.1). In fact, many verbs end in a vowel and almost all nouns start with a vowel, which is why transitive verbs fuse with their nominal object through vowel deletion (see Example 3.149).

(3.149) (a) *lɔ ísaj → [lísaj]*
throw stone
'throw a stone' (ibe203-00.157)

(b) *to idʒej → [tidʒej]*
pound pounded_yam
'pound yam' (ibe231-00.121)

In many cases, the verbal vowel and tone is deleted and the only verbal trace is the initial consonant. The underlying verbal form only appears in cases where the object does not immediately follow the verb, such as in content questions (Example 3.150).

(3.150) (a) *màd ɔ́ lɔ? → [màdɔ́ lɔ]*
what 3SG.HUM throw
'what did s/he throw?' (ibe203-00.154)

(b) *màd ɔ̀ é to? → [màdɔ́ˈɔ́ é to]*
what 2SG PROG pound
'what are you pounding?' (ibe001-02.167)

However, some verb-noun combinations are not separable, as they only appear in combination and each part's meaning is impossible to disassemble. In some cases, only the noun can be retrieved, such as in *jotu* 'steal', where the noun *otu* 'thief' is recognizable, but the verb cannot be isolated.

Many CVCV verbs might have originated from verb-noun combinations that are not separable anymore. This is probably the reason for long verbs having non-matching prefix and suffix-VH. As the underlying verb has a different ATR value than the noun, their prefix vowels differ from their suffix vowels (consult Section 3.1.4.4 for examples).

3.6.4 Reduplication

Full reduplication is used frequently in Abesabesi and has different functions. Reduplicated nouns could be seen as compounds and adapt a new or modified meaning. *oŋo* 'wife' becomes *oŋóoŋo* 'woman', *ɔban* 'child' becomes *ɔbánɔban* 'small child', or *ɔni* 'person' becomes *ɔnɔ́ɔni* 'everybody'. Reduplications of regular nouns require the GenH, but instead of deleting the first part's final vowel, the second part's initial vowel is assimilated (e.g., *ègbo* + *ègbo* > *ègboògbo* 'first born'). The only exception is when the first part's final vowel is /i/ (e.g., *ɔni* + *ɔni* > *ɔnɔ́ɔni* 'everybody'). In this case, the first part's final vowel is deleted.

3.6.4.1 Property nouns

Reduplicated property nouns do not carry a GenH and undergo a V_2 deletion. E.g., *ɛŋa* + *ɛŋa* > *ɛŋaŋa* 'new', *efo* + *efo* > *efofo* 'white', or *ɛmiŋɛ* + *ɛmiŋɛ* > *ɛmiŋɛmiŋɛ* 'small'. Property nouns do not change meaning when reduplicated. Some tend to be reduplicated when used in the fronted focus position, as an adverbial, or as predicate nominal. Example 3.151 a shows a reduplicated property noun in the fronted focus position, Example 3.151 b as an adverbial and Example 3.151 c as a predicate nominal. Non-reduplicated property nouns in these positions are rather rare, such as in Example 3.151 d.

(3.151) (a) *efofo gí é tu èhen*
white FOC 3SG.NHUM be now
'it is white now' (ibe178-02.120)

(b) *sá'á ʃiga èŋèmèŋèm*
2SG.Q get_up well
'did you get up well?' (ibe140-00.216)

(c) *ɔmìŋà tu ɛŋaŋa è*
knife be new NEG
'the knife is not new' (ibe365-00.176)

(d) *ɛbuj ìnɛ́ tu efo?*
goat how_many be white
'how many goats are white?' (ibe033-00.134)

3.6.4.2 Numerals

Numerals adapt a distributive meaning when reduplicated. This is a fairly common strategy in the wider region and is also present in Yoruba (Sachnine, 2014, p. 68) and Nigerian English.

(3.152) *bà* *á* *ba* *ekìn* *ekìn*
3PL.HUM PROG come one one
'they are coming one by one' (ibe059-00.017)

3.6.4.3 Non-specific indefinite (pro)nouns

Abesabesi creates non-specific indefinite nouns according to the schema N + GenH + t + N. These nouns translate as 'any X', 'whichever X' or, in negated clauses, 'no X'. *Okpo* 'way', for example becomes *okpótokpo* 'any/whichever/no way', and *iʃo* 'house' becomes *iʃótiʃo* 'any/whichever/no house'.

Non-specific indefinite pronouns are created from the abstract nouns *ɔni* 'person', *iniŋ* 'thing', *òjè* 'place', and the numeral *ekìn* 'one'. They use a slightly different schema. Only the first vowel of the base noun is reduplicated and *ńt* is put between the reduplicated vowel and the noun (Table 3.60). Additionally, the reduplicated forms of *ɔni* 'person' and *iniŋ* 'thing' undergo further idiosyncratic changes: *ɔńtɔn* 'anybody/nobody/whoever' and *ińtiǹ* 'anything/nothing/whatever'. Section 3.5.2.9 discusses their function.

Base noun	Meaning	Indefinite pronoun	Meaning
ɔni	person	ɔńtɔn	anybody, nobody, whoever
iniŋ	thing	ińtiǹ	anything, nothing, whatever
ekìn	one	eńtekìn	anything, nothing, whatever
òjè	place	òńtòjè	anywhere, nowhere, wherever

Table 3.60: Non-specific indefinite pronouns

3.6.4.4 Non-possessive reduplication

The non-possessive reduplication is a phenomenon which is also present in Yoruba and might be a calque. In Abesabesi, it has the schema N + *ani* 'people' + N. The whole construction indicates that something (the noun N) does not belong to the person in question and could be translated as 'other people's N'.

(3.153) (a) *bà* *gbà* *àmùn* *ani* *àmùn* *bèe* *mĩ* *àmùn*
3PL.HUM put language people language to do language
'they use another people's language to speak' (ibe140-00.123)

(b) *náà* *tu* *ɔban* *ani* *ɔban* *ʃa* *áje* *no*
1SG.NEG be child people child for mother 1SG.POSS
'I am not a stranger's child to my mother' (ibe140-00.053)

3.7 Clause structure

This section expands the overview of the sentence given in Section 3.2 and discusses further aspects of the clause structure in Abesabesi. Section 3.7.1 describes how arguments are marked, Section 3.7.2 deals with the structure of predicate nominals and related constructions, Section 3.7.3 discusses the structure and position of adverbs, and Section 3.7.4 presents the structure of embedded clauses.

3.7.1 Argument structure

As Abesabesi does not have a case marking system for core arguments, subject and object marking is based on word order. The preverbal argument position is reserved for the subject, while the postverbal position is reserved for one or two objects. In clauses with an intransitive verb, the subject is the only argument and can have an agent-like (Example 3.154 a) or a patient-like (Example 3.154 b) role.

(3.154) (a) *èɡèdʒi á ʃu*
horse PROG run
'a horse is running' (ibe224-00.091)

(b) *ɛɲɔ hu*
snake die
'the snake died' (ibe224-00.029)

In clauses with transitive verbs, the subject is the more agent-like argument, while the object is the more patient-like argument.

(3.155) *oɲo no bug ɛbuj*
wife 1SG.POSS beat goat
'my wife beat the goat' (ibe327-00.115)

Verbs for bodily states and feelings are a special case. Such verbs are called "active body state verbs" ABSV henceforth. From their syntactic structure, these body states or feelings are active instances that superimpose the feeling onto a patient. A noun for the body state or feeling (or in some cases a non-human pronoun) occupies the subject position of an ABSV and the experiencer occupies the object position.

(3.156) (a) *imu á kɛ̀m na*
hunger PROG ABSV_hunger 1SG.OBJ
'I am hungry' (ibe017-00.105)

(b) *ɛ́ dìg no*
3SG.NHUM.SBJ ABSV_sickness 1SG.OBJ
'I am sick' (ibe182-00.112)

(c) *ɔ̀ʃɪ̀ŋ é mi no*
cold PROG ABSV_cold 1SG.OBJ
'I am feeling cold' (ibe224-00.031)

In clauses with ditransitive verbs, the position immediately after the verb is reserved for the indirect object and the second postverbal position is reserved for the direct object. Abesabesi is thus a primary object language in the sense of Dryer (1986), in that the indirect object of a ditransitive verb holds the same position as a direct object of a transitive clause. Therefore, instead of distinguishing direct and indirect objects, the terms "primary" and" secondary" objects will be used. Any object directly following the verb will be called "primary object" (PO) and objects in the second postverbal position will be called "secondary object" (SO). Example 3.157 shows two utterances with ditransitive verbs. In Example 3.157 a, for instance, the recipient *ès* '1PL.OBJ' immediately follows the verb and is thus called the primary object. The theme *ɛmũ* 'money' holds the second postverbal position and is thus the secondary object.

(3.157) (a) *bà* *ki* *ès* *ɛmũ*
3PL.HUM.SBJ give 1PL.OBJ money
'They gave us money' (ibe057-00.198)

(b) *bà* *kɔ́* *na* *èkirɔ̀m*
3PL.HUM.SBJ teach 1SG.OBJ Ekiromi
'they taught me Ekiromi' (ibe089-00.025)

Other optional arguments are added through participant-adding auxiliaries. These auxiliaries take primary objects and mark a particular thematic role. Their position can be pre- (Example 3.158 a) or postverbal (Example 3.158 b).

(3.158) (a) *bà* *ton* *no* *á* *kɔn*
3PL meet 1SG.OBJ PROG fight
'they fight against me' (ibe057-00.159)

(b) *bà* *nàbù* *kpa* *na*
3PL.HUM.SBJ wait for 1SG.OBJ
'they waited for me' (ibe057-00.188)

What has been called the dative object is in fact also the primary object of the auxiliary *ʃa*. However, dative objects differ from other auxiliary verb constructions on two points. Firstly, it is obligatory in clauses with the transitive verb *gbà* 'put', which are equivalent to clauses with the ditransitive verb *ki* 'give'.

(3.159) (a) *bà* *gbà* *ìwé* *sá*
3PL.HUM put book 2SG.DAT
'they gave you a book' (ibe089-00.010)

(b) *bà* *gbà* *ìwé* *ʃa* *ɔmis*
3PL.HUM put book DAT king
'they gave a book to the king' (ibe089-00.016)

And secondly, its pronominalized form has formed an incomplete paradigm set through contraction (see Section 3.5.2.6).

Subjects and objects are also distinguished in pronouns. There are two different paradigms for subject and object pronouns (see Section 3.5.2.1 and 3.5.2.2). Object pronouns do not distinguish primary from secondary objects. If the primary and secondary object are both pronominalized, they are distinguished through word order with the secondary object following the primary object.

(3.160)

mi	*ɔ́*	*ba*	*haǹ*	*ekìn*	*ìwùd*	*òroǹbó*	*ìdɛ́n*	*si*	*ba*	*ki*
SIP	3SG.HUM	PST	pick	one	PAR	orange	those	SEQU	PST	give
u		*ɛ*								
3SG.HUM.OBJ		3SG.NHUM.OBJ								

'he picked one out of the oranges and gave it to him' (ibe229-00.096)

3.7.2 Predicate nominal clauses and related constructions

This section discusses those constructions Payne (1997, pp. 111-128) titles "predicate nominals and related constructions": predicate nominal clauses, attributive clauses, predicate locative clauses, existential clauses, and possessive clauses. Abesabesi uses the copula verbs *tu*, *wa*, and *dʒe*, the defunct verb *té*, and the regular verb *ʃi* 'have' for these constructions (see Table 3.61)

Clause type	Verb
Predicate nominal clauses (default)	*tu*
Predicate nominal clauses (hypothetical)	*wa*
Predicate nominals clauses (names/titles)	*dʒe*
Attributive clauses (property nouns)	*tu*
Attributive clauses (property verbs)	–
Predicate locative clauses	*tu*
Existential clauses	*té*
Possessive clauses	*ʃi*

Table 3.61: Verbs used in predicate nominals and related constructions

Proper inclusion and equation are not differentiated in Abesabesi, but three types of predicate nominal clauses can be distinguished. The default case is the use of the copula verb *tu*. In affirmative clauses with *tu*, the predicate nominal is usually focus fronted (Section 3.8.1). Example 3.161 shows a proper inclusion (a) and an equation (b) with the copula verb *tu*.

(3.161) (a)

ɔban	*gí*	*ɔ̀*	*tu*
child	FOC	2SG	be

'you are a child' (ibe015-00.005)

(b)

áje	*no*	*gí*	*ɔ́*	*tu*
mother	1SG.POSS	FOC	3SG.HUM	be

'she is my mother' (ibe015-00.027)

If the equation or proper inclusion is hypothetical, e.g. the clause is a counterfactual or hypothetical conditional, or is situated in the future, the copula *wa* is used. Example 3.162 shows one equation in the future (a), a hypothetical conditional (b), and a counterfactual conditional (c).

(3.162) (a) *owos so gí ǹ bá wa*
husband 2SG.PROG FOC 1SG FUT be
'I will be your husband' (ibe015-00.063)

(b) *ɛ́ˡɛ́ wa ɔ̀kɔdɔ́ɔni hu*
3SG.NHUM.COND be old_person die
'if an old person dies' (ibe059-00.002)

(c) *ɛ́ˡɛ́ wa kpé ǹ ka kàj ìwé dáadáa,*
3SG.NHUM.COND be COMP 1SG PRF read book well
naá màá páàs
1SG.HAB HAB pass
'if I had studied properly, I would have passed' (ibe168-00.070-071)

In equations where the predicate nominal is a name or title the copula verb *ʤe* is used (Example 3.163).

(3.163) (a) *ònì gí ʤe júnis*
1SG.IDP FOC be <NAME>
'I am Jonas' (ibe017-00.028)

(b) *ìmùn oɲo u ʤe ɲánípo*
name wife 3SG.HUM.POSS be <NAME>
'his wife's name is Yannibo' (ibe105-00.006)

Attributive clauses are clauses where a property is attributed to a referent. As properties can be encoded in property nouns and property verbs, these two cases have to be distinguished. Attributive clauses with property nouns use the copula verb *tu* to attribute the property of the property noun to the subject. In predicative use, property nouns often appear in the reduplicated form (Section 3.6.4). Example 3.164 a shows an utterance where the property noun *ɛŋa* 'new' is reduplicated. As the copula *tu* is a verb, it can take TAM markers such as the HAB marker *màá/mèé/mòó*. Attributive clauses with property verbs do not require a copula verb because the property verb functions as a verb (Example 3.164 b).

(3.164) (a) *ìwé màá tu ɛŋaŋa*
book HAB be new
'the book used to be new' (ibe363-00.160)

(b) *ìsɔ̀m èéni ɲà*
soup this be_red
'this soup is red' (ibe147-00.162)

Predicate locative clauses require the copula *tu* and the predicative locative must be marked with the LocH.

(3.165) *ìkpàr* *ba* *tu* *H* *ìdʒɔ̀*
children 3PL.HUM.POSS be LocH farm
'their children are at the farm' (ibe005-00.300)

Existential clauses employ the defunct verb *té*. They are discussed in Section 3.4.4.3.

Possessive clauses use the verb *ʃi* 'have'. While the possessor is the subject, the possessed is the object of *ʃi*.

(3.166) *ɔ́* *ʃi* *ìdʒɔ̀*
3SG.HUM have farm
's/he has a farm' (ibe005-00.280)

The negation of the clause types discussed in this section will be discussed in Section 3.7.6.2.

3.7.3 Adjuncts

Adjuncts are located right after the VPs or in the fronted focus position. The only exception to that is the auxiliary verb *gɛn*, which appears even after the adjuncts (see Example 3.167). An adjunct can consist of a PP (Example 3.168 a), an adverb (Example 3.168 b), or any noun denoting a location with the LocH (Example 3.168 c)[14].

(3.167) *TAM* *MainV* *PO* *ADJUNCT* *PostV*
sí wuʃ i lí atɔ gɛn
SEQU gather 3PL.NHUM LOC ground finish
'after he brings them out to the ground' (ibe023-00.020)

(3.168) (a) *PRO* *V* *PO* *ADJUNCT*
ìbùrɛ́ mi ɔ́ ɲì onu ʃí atɔ
time REL 3SG.HUM open mouth LOC ground
'when he opened his mouth to the ground' (ibe111-00.039)

(b) *PRO* *TAM* *V* *ADJUNCT*
à bé mĩsum̀ èʃɔj
1PL FUT work tomorrow
'we will work tomorrow' (ibe190-00.079)

(c) *MainV* *PO* *PostV* *PO* *ADJUNCT*
oòsì, mĩ iniŋ èɲèm ʃa ès H iʃo
God do thing good DAT 1PL.OBJ LocH house
'God, do us a favor in the house!' (ibe052-00.062)

Adjuncts are optional parts of the sentence that have an adverbial function. They contextualize the event, typically by adding a time, location, or manner.

[14]Note that the LocH is usually not marked, such as in prior examples like Example 3.103 b, which is repeated here as Example 3.168 c, but with the LocH marked

3.7.4 Subordination

Embedded clauses are part of a sentence but depend on a higher-level clause – the "matrix clause". The highest level clause, which is independent of other clauses, is called the main clause. Abesabesi embedded clauses differ from main clauses, as they are introduced by specific markers. Embedded clauses that follow the main clause do not contain the 1H. Abesabesi exhibits eight types of embedded clauses that can formally be distinguished. They can perform various functions depending on the context and on where they appear within the sentence.

Name	Introduced with	Subsection
Complement clause	Complementizer *mí* or *kpé*	3.7.4.1
Relative clause	RelH and the relativizer *mi*	3.7.4.2
Conditional clause	HLH melody	3.7.4.3
Causal clause	*torí kpé*	3.7.4.4
Causal/final clause	*í tɔ̀n kpé*	3.7.4.4
?	*àba/àbe/àbo*	3.7.4.4
?	*bàa/bèe/bòo*	3.7.4.4
?	*sá wá/sé wé/(só wó?)*	3.7.4.4

Table 3.62: Embedded clause types

Complement clauses, relative clauses, and conditional clauses are frequently used clause types that are each discussed in their respective subsections. The final subsection will present the other five minor clause types that are used infrequently in the corpus and could thus not be sufficiently analyzed.

3.7.4.1 Complement clause

Complement clauses are embedded clauses that are used as arguments of complement-taking verbs. Many complement-taking verbs are verbs of speech, knowledge, control, or desire: *mò* 'say', *ʃìʃa* 'tell', *bìj* 'ask', *sà* 'know', *jas* 'agree', *ʃì* 'let', *ʃeʃè* 'be possible', and *dò* 'want'. Some complement-taking verbs are loan words from Yoruba: *rò* 'think', or *kábàámɔ̀* 'regret'.

(3.169) (a) *ǹ sà kpé ijes aje no bá jagas*
1SG know COMP body mother 1SG.POSS FUT get_well
'I know my mother will get well' (ibe043-00.117)

(b) *ɛ́ ʃeʃè mí abíɔ́dű́ wa pátì*
3SG.NHUM be_possible COMP <NAME> be party
'It is possible that Abiodun was at the party' (ibe281-00.411)

(c) *ɔmis kpàʃe kpé mí ǹ ʃìga ègur*
king order COMP COMP 1SG leave town
'the king ordered that I should leave the town' (ibe043-00.040)

Abesabesi has two complementizers, *mí* and the Yoruba loan *kpé*. In many cases, they appear to be interchangeable and, in some cases, they are used together (Example 3.169 c).

Complement clauses follow the complement-taking verb and their structure does not differ much from that of main clauses. The only differences are the use of a complementizer to introduce complement clauses and the use of logophoric pronouns whenever the subject referents of the matrix clause and the complement clause are the same. Logophoric pronouns and their use are discussed in Section 3.5.2.5.

(3.170) *ɔ́ɔ̀ dò mí àu miʃ è*
3SG.NEG want COMP 3SG.HUM.LOG sleep NEG
's/he does not want to sleep' (ibe346-00.235)

Complement clauses have a second function. When used with one of the temporal auxiliaries *wɛǹ* 'do before' (Example 3.171 a) or *gɛn* 'do after' (Example 3.171 b), they exhibit a temporal meaning.

(3.171) (a) *ɔ́ ka das mí ǹ wɛǹ niŋ iʃo*
3SG.HUM PRF go COMP 1SG do_before arrive home
's/he left before I arrived at home' (ibe190-00.151)

(b) *mí ɔ́ wuʃ i lɔ gɛn*
COMP 3SG.HUM collect 3PL.NHUM throw do_after
'after he threw them inside' (ibe023-00.031)

3.7.4.2 Relative clause

A relative clause modifies a particular noun or pronoun of the matrix clause. Following the relativized NP, it consists of the RelH on the right end of the relativized NP, the relativizer *mi*, and a clause.

(3.172) *ɔ́ˡɔ́ je ɔhunɛ H mi tu atɔ?*
2SG.Q see tree RelH REL be ground?
'did you see the tree that is on the ground?' (ibe292-00.406)

The relativizer *mi* has various allomorphs depending on the segment that follows. Before most consonants, it behaves like a syllabic nasal, assimilating with the consonant's place of articulation. Before syllabic nasals, it is produced as [mi]. Before vowels, the /i/ is deleted and it is produced as [m].

(3.173) (a) *wol ɛbuj H mi ton ɛsagum á kɔn*
kill goat REL meet sheep PROG fight

→ [wol ɛbʷijínton ɛsagum ákɔn]
'kill the goat that is fighting with the sheep' (ibe108-00.038)

(b) *ɔni H mi ǹ bá jen gí tu èéni*
person RelH REL 1SG FUT marry FOC be this

→ [ɔní mi᷅bájen gí tèén]
'this is the person I will marry' (ibe108-00.104)

(c) *àn bá kàj ìwé H mi àn de èʃì*
2PL FUT read book RelH REL 2PL buy market

→ [àmbákàj ìwé màn déʃì]
'you will read the book you bought at the market' (ibe108-00.068)

Relative clauses are nominal modifiers and thus part of an NP. Their position within an NP is discussed in Section 3.3.4.5. As Abebesabesi employs a gapping strategy, the canonical word order within the relative clause may lack an obligatory argument slot. In this case, the slot is filled by the relativized NP. The syntactic role of the relativized NP can easily be derived from the gap within the relative clause as Abesabesi's word order is SVO. Within the relative clause, relativized NPs can fulfill the following syntactic functions: subject (Example 3.174 a), primary object (Example 3.174 b), secondary object (Example 3.174 c), primary object of auxiliary verbs (Example 3.174 d, e), adverbial (Example 3.174 f), and attribute (Example 3.174 g). If it has the function of an attribute, a resumptive pronoun has to be used (Example 3.174 g). In all other cases, resumptive pronouns are not used.

(3.174) (a) *ǹ dɛ̀n ɔni mi sɛmɛ̀ na*
1SG like person REL greet 1SG.OBJ
'I like the person that greeted me' (ibe208-00.002)

(b) *ǹ dɛ̀n ɔni mi fű́mi jen*
1SG like person REL <NAME> marry
'I like the person Funmi married' (ibe208-00.007)

(c) *ǹ dɛ̀n àmùn mi ɔ́ kɔ́ na*
1SG like language REL 3SG.HUM teach 1SG.OBJ
'I like the language s/he taught me' (ibe208-00.015)

(d) *ǹ dɛ̀n ɔni mi fű́mi ʃag atɔ ʃa*
1SG like person REL <NAME> sweep floor DAT
'I like the person Funmi swept the floor for' (ibe208-00.029)

(e) *ǹ dɛ̀n ɔni mi ǹ ton kɔn ijɛ̀n*
1SG like person REL 1SG meet fight yesterday
'I like the person I fought with yesterday' (ibe208-00.036)

(f) *ǹ dɛ̀n ègur mi bò bo dʒo so*
1SG like town REL 3PL.HUM PST give_birth 2SG.OBJ
'I like the town where you were born' (ibe208-00.060)

(g) *ǹ dɛ̀n ɔni mi iʃo u ga fàŋ*
1SG like person REL house 3SG.HUM be_big CPOR/SUP
'I like the person whose house is the biggest' (ibe208-00.068)

Conjunctions and pronouns that introduce specific types of adverbial or relative clauses, as the ones in many European languages, do not exist in Abesabesi per se. However, generic nouns and functional nouns such as *iniŋ* 'thing', *ɔni* 'person', *ìbùrɛ* 'time', *òjè* 'place' or *èdʒì/èdʒèdʒì* 'like' are modified by relative clauses to fulfil the same function as conjunctions like 'what', 'who', 'when', 'where' or 'as'. This way, relative clauses can function as complement, locative, temporal, or manner clauses.

(3.175) (a) *ʃiʃa na iniŋ mi ɔ̀ bé mĩ idédumon*
tell 1SG.OBJ thing REL 2SG FUT do evening
'tell me what you will prepare for dinner' (ibe108-00.121)

(b) *fűmi jɔbɔ bàá jen ɔni mi ɔ́ ʃen*
<NAME> can ? marry person REL 3SG.HUM be_good
u ajo èdɛ́n
3SG.OBJ eye that
'Funmi may marry who she likes' (ibe117-00.010)

(c) *ʃiʃa na ìbùrɛ́ mi à bá das*
tell 1SG.OBJ time REL 1PL go
'tell me when we will leave' (ibe108-00.156)

(d) *gbà òjè mi ɔ̀ tu titì na*
put place REL 2SG be show 1SG.OBJ
'show me where you live.' (ibe108-00.138)

(e) *ǹ dɛ̀n èdʒí mi bè mĩ ʃe bós ìsɔ̀m*
1SG like how REL 3PL.HUM do do_how cook soup
'I like how they cook the soup' (ibe108-00.195)

3.7.4.3 Conditional clause

A conditional clause expresses a condition about its matrix clause, in which the consequence is expressed. Abesabesi does not mark the consequence clause specifically, but the conditional clause is marked by a HLH melody on the lengthened last vowel of the subject phrase. As the lengthened vowel can only carry two tones, this results in two high tones with a downstep in-between.

(3.176) *eji jel ìdʒɛ, náà bá fù ìguj̀ è*
rain rain today 1SG.NEG FUT wash clothes NEG

→ [ejíˈí jel ìdʒɛ náàbáfùguj̀è]
'if it rains today, I will not wash clothes' (ibe168-00.004)

For pronominal subjects, the conditional pronoun paradigm is used, which is composed of the irrealis pronouns and the conditional HLH melody (Section 3.5.2.1).

(3.177) *náˈá toǹ ìwé no ba ìdʒɛ, ǹ bá kàj*
1SG.COND carry book 1SG.POSS come today 1SG FUT read
ɛ
3SG.NHUM.OBJ
'if I bring my book today, I will read it' (ibe168-00.038)

The conditional clause is also used for temporal clauses that express immediate posteriority. This can be seen in the two consecutive utterances of Example 3.178.

(3.178) (a) *bá$^{\downarrow}$á sù àlàj ʃí ewùlò èdɛ́n gɛn*
3PL.HUM.COND stick yam_stake LOC yam that finish
'after they stick the yam_stakes by the yam plants' (ibe035-00.020)

(b) *mi ewùlò ɛ bá ŋanà*
SIP yam the FUT geminate
'the yam will geminate' (ibe035-00.021)

Negative conditionals are introduced by *m̀* and lack the HLH melody. They are discussed in Section 3.7.6.2.

Most conditional sentences have the order conditional clause > consequence clause. Conditional clauses that follow the consequence clause start with the marker *sàm* and do not feature the HLH melody.

(3.179) *ɔni gbɛ́dɔ̀ haǹ étù fe, sàm ɔ́ á tu iʃum*
person must pick hat wear COND 3SG.HUM PROG be on
ɔ̀kadà
motor_taxi
'one must wear a helmet when using a motor taxi' (ibe111-00.002)

Counterfactual clauses are a special kind of conditional clause where the condition cannot be or is unlikely to be fulfilled. Abesabesi has a construction to express counterfactuals, consisting of the conditional 3SG.NHUM pronoun, the hypothetical copula verb *wa*, and a complement clause introduced with *kpé*. This construction could literally be translated as 'if it was that...'.

(3.180) (a) *ɛ́$^{\downarrow}$ɛ́ wa kpé bà ka fù òli*
3SG.NHUM.COND be COMP 3PL.HUM PRF wash clothes
ìkúdùkud, ií mòó ko hu èhen
morning 3PL.NHUM.HAB HAB PRF dry now
'if they had washed the clothes in the morning, they would be dry by now' (ibe168-00.054)

(b) *ɛ́$^{\downarrow}$ɛ́ wa kpé sáà tu èkó*
3SG.NHUM.COND be COMP 2SG.NEG be <PLACE-NAME>
ijɛ̀n è, neé mèé je so
yesterday NEG 1SG.HAB HAB see 2SG.OBJ
'if you had not been in Lagos, I would have seen you' (ibe168-00.073)

(c) *ɛ́$^{\downarrow}$ɛ́ wa kpé ǹ sà kpé ní ɔ̀ á ba,*
3SG.NHUM.COND be COMP 1SG know COMP 2SG PROG come
ǹ ka bá kèd so èkó
1SG PRF FUT carry 2SG.OBJ <PLACE-NAME>
'if I had known you were coming, I would have picked you up in Lagos' (ibe281-00.136-137)

3.7.4.4 Other clause types

Abesabesi also exhibits a few minor clause types which only have a few occurrences in the corpus. They are merely presented here, as their analysis is incomplete.

There are three non-finite constructions[15] that seem to express purpose or are used as complement clauses. One construction is introduced with *àba/àbe/àbo*[16].

(3.181) *ìwùd na ɲum̀ abò nunù so*
stomach 1SG be_sweet CONJ meet 2SG.OBJ
'I am happy to meet you' (ibe017-00.042)

Another construction is introduced with *bàa/bèe/bòo*.

(3.182) *ɔ́ ka das bèe go*
3SG.HUM PRF go CONJ pray
'He has gone to pray' (ibe331-00.232)

And a third construction is introduced with *sá wá/sé wé* (and probably also *só wó*). Example 3.183 contains three consecutive clauses. The main clause (a) is followed by two clauses introduced with *sá wá/sé wé* (b, c).

(3.183) (a) *mi àba ídiàn ba wa èle òhun*
SIP 3PL.HUM.IDP two PST stay home here
'two of them stayed home' (ibe065-00.080)

(b) *sé wé mĩ isum̀ idʒɔ̀*
CONJ do work farm
'to work on the farm' (ibe065-00.081)

(c) *sá wá tɔ́dʒú àbès ìkpàr*
CONJ take_care 1PL.IDP
'and to take care of us children' (ibe065-00.082)

Apart from these non-finite clauses, there are two more finite clause types. Firstly, the conjunction *torí kpé* is borrowed from Yoruba and expresses causality.

(3.184) *isum̀ kpɔ̀ è, torí kpé isum̀ tu ìdʒɔ̀ è*
work be_much NEG because work be farm NEG
'there is not much work because there is no work on the farm'
(ibe167-00.017+018)

The Abesabesi equivalent is the causal construction *í tɔ̀n kpé*.

(3.185) *náà ba iʃo è í tɔ̀n kpé sáà*
1SG.NEG come home NEG 3PL.NHUM be_caused_by COMP 2SG.NEG
sàg na è
call 1SG.OBJ NEG
'I did not come home because you did not call me' (ibe265-00.319)

[15]Non-finite clauses in Abesabesi have no subject nor TAM markers

[16]The underlying tones on this word could not be analyzed. That is why the tones in the transcription differ.

The verb *tɔ̀n* seems to mean 'to cause' and 'to be caused by'. The construction *í tɔ̀n kpé* is thus also used as a final clause. This is illustrated by example 3.186. While the first utterance (a) contains the main clause, the second utterance (b) contains the final clause introduced with the negative form of *í tɔ̀n kpé* (*ûi tɔ̀n kpé*). This example contains a long form of the Yoruba complementizer (*kpé ní*) together with the Abesabesi complementizer *mí*.

(3.186) (a) *bò bó hùr òkpònò, bò hùr i ʃa*
3PL.HUM FUT put_into cotton 3PL.HUM put_into 3PL.OBJ DAT
ahuǹ, bò hùr i ʃa asug
nose 3PL.HUM put_into 3PL.OBJ DAT ear
'they will insert cotton, they will put it into the nose and into the ear' (ibe059-00.074)

(b) *ûi tɔ̀n kpé ní mí ɔ́taj̀ u ba*
3PL.NHUM.NEG cause COMP COMP smell 3SG.HUM.POSS PST
bùr ani ʃí ɔ̀ʃum è
chase people LOC back NEG
'so that its smell would not chase people away' (ibe059-00.075)

3.7.5 Questions

This section discusses various interrogative constructions in Abesabesi which are used to ask questions. We distinguish polar questions (Section 3.7.5.1) from content questions (Section 3.7.5.2). While polar questions (yes-no questions) ask about the truth value of a proposition, content questions (wh-questions) ask about a specific piece of information using interrogative pronouns. Both types of questions can be embedded, which is described in the last subsection (3.7.5.3).

Common to all question types, except for embedded content questions, is the sentence-final question particle *áj̀*. It is optional in independent questions but appears to be obligatory in embedded polar questions.

3.7.5.1 Polar questions

Polar questions are indicated by a sentence-initial high tone – the QH. It attaches to the first mora of the subject (Example 3.187 a, b). If the subject starts with a non-syllabic consonant, an initial /i/ is pronounced in order to carry the QH (see Example 3.187 c)

(3.187) (a) *H ebo H é dʒe iniŋ? → [ébóédʒiniŋ]*
QH dog 1H PROG eat thing
'is the dog eating?' (ibe363-00.115)

(b) *H ìwé H é dʒida? → [íʲwéédʒida]*
QH book 1H PROG burn
'is the house burning?' (ibe363-00.120)

(c) *H délé H á gbà ewùlò taŋ?*
QH <NAME> 1H PROG put yam soak

→ [ídélé ágbewùlò taŋ]
'did Dele soak the yam?' (ibe345-00.036)

The QH also attaches to pronouns. The resulting paradigm, the interrogative pronouns, is a combination of bound subject pronouns, QH, and 1H. They are introduced in Section 3.5.2.1. Example 3.188 contains the interrogative pronoun *ɔ́ˡɔ́*, which is a combination of the QH, the bound subject pronoun ɔ̀ '2SG.SBJ', and the 1H.

(3.188) *ɔ́ˡɔ́ ka hɔg?*
2SG.Q PRF wash
'have you already taken your bath?' (ibe331-00.339)

Alternative to the QH, but very infrequently, the particle *(mi)ní* is used to express polar questions. This particle might be the origin of the QH. Example 3.189 shows the three options to mark a polar question (QH, ní, and miní).

(3.189) (a) *H ɔ̀ H deŋ? → [ɔ́ˡɔ́deŋ]*
QH 2SG.SBJ 1H climb
'did you climb?' (ibe363-00.115)

(b) *ní ɔ̀ H á miʃ? → [nɔ́ˡɔ́ámiʃ]*
Q 2SG.SBJ PROG sleep
'are you sleeping?' (ibe363-00.120)

(c) *miní ɔ̀ H deŋ → [minɔ́ˡɔ́deŋ]*
Q 2SG.SBJ 1H climb
'did you climb?' (ibe345-00.036)

The question particle *áj̀* is optional in independent polar questions (Example 3.190).

(3.190) (a) *ɔ́ˡɔ́ ke dʒiniŋ?*
2SG.Q PRF eat
'have you already eaten?' (ibe017-00.145)

(b) *ɔ́ˡɔ́ ke dʒiniŋ áj̀?*
2SG.Q PRF eat Q
'have you already eaten?' (ibe043-00.037)

Negative polar questions also employ the QH or the particle *(mo)ní*, which is the negative equivalent of *miní*. Otherwise, they adhere to the rules of negated sentences: no 1H, the use of irrealis pronouns, and the negation particle *è* (see Section 3.7.6).

(3.191) (a) *H sáà kpas na è?*
QH 2SG.NEG hear 1SG.OBJ NEG
'don't you hear me?' (ibe186-00.145)

(b) *ní náà sɛmɛ̀ sa è?*
Q 1SG.NEG greet 2SG.OBJ NEG
'did I not greet you?' (ibe265-00.267)

(c) *moní ɔ̀ʃɪ̀ŋ é mi so è?*
Q.NEG cold PROG ABSV_cold 2SG.OBJ NEG
'are you not cold?' (ibe265-00.258)

3.7.5.2 Content questions

Content questions contain a fronted interrogative phrase and the rest of the clause following that phrase. The interrogative phrase may consist of an interrogative pronoun (Section 3.5.2.12) or an NP modified by an interrogative pronoun.

(3.192) (a) *máàn ɔ̀ gbòm?*
who 2SG.SBJ hit
'who did you hit?' (ibe246-00.208)

(b) *ani mèj tu ìdín?*
people where be these
'who are these people? (lit.: these are people from where?)'
(ibe108-00.102)

The position of the interrogative phrase is always the first position of the sentence. The 1H therefore attaches to the right end of it.

(3.193) *ɔban máàn H ɔ̀ tu → [ɔbánmáànɔ́ɔ̀tu]*
child who 1H 2SG.SBJ be
'whose child are you?' (ibe017-00.032)

The question particle *áj̀* in content questions only appears rarely in the corpus. All instances occurred in direct or indirect speech.

(3.194) (a) *màd ǹ bé be de áj̀*
what 1SG.SBJ FUT PST buy Q
'what should I buy?' (ibe105-00.042)

(b) *ani gbà ès mijɛ̀jɛ̀ kpé isuǹ ebòn ɔ́ kɔ́ ès áj̀*
people put 1PL.OBJ mock COMP work which 3SG teach 1PL Q
'people mocked us [asking us] which work he had taught us'
(ibe065-00.150)

3.7.5.3 Embedded questions

Some questions are not independent, but function as a complement of certain verbs of speech or knowledge. These embedded indirect questions can be polar or content questions.

(3.195) (a) *ɔ́ bij no kpé ǹ dɛ̀n ewùlò áj̀*
3SG ask 1SG.OBJ COMP 1SG like yam Q
's/he asked me whether I like yam' (ibe281-00.321)

(b) *náà sì tam iniŋ mi kpɛ̀ è*
1SG.NEG anymore remember thing REL remain NEG
'I don't remember any more what remains' (ibe023-00.098)

(c) *ǹ kɔ́ èdʒí mi bè mĩ́ é to idʒej*
1SG learn as REL 3PL.HUM do PROG pound pounded_yam
'I learned how they make pounded yam' (ibe043-00.083)

As seen in the examples, embedded questions do not employ the QH nor the interrogative pronouns. Instead, the default pronoun paradigm is used.

While embedded polar questions always appear with the question particle *áj̀* (Example 3.195 a), embedded content questions do not appear with it. Furthermore, they do not employ interrogative pronouns but abstract nouns modified by relative clauses. 'What', is expressed through *iniŋ mi* 'the thing which', 'who' through *ɔni mi* 'the person who', 'when' through *ìbùrɛ mi* 'the time that', 'where' through *òjè mi* 'the place where', and 'how' through *èdʒì mi* 'the way that' (Examples 3.195 b, c).

3.7.6 Negation

Negation in Abesabesi is mainly expressed through the clause-final clitic *è*, which appears in all negated clauses.

(3.196) (a) *ìwé bá lògìnò*
book FUT break
'the book will break' (ibe363-00.036)

(b) *ìwé bá lògìnò è*
book FUT break NEG
'the book will not break' (ibe363-00.192)

Phonologically, the particle behaves similarly to the progressive marker *á/é/ó* (see Section 3.4.2.2). It is pronounced [jè] when preceded by a consonant or a high vowel (/i/ or /u/). When preceded by any other vowel, it is pronounced [è]. Sometimes, an epenthetic /i/ is added when a *è* follows a consonant.

(3.197) (a) *iʃo no ga è → [iʃóno gaè]*
house 1SG.POSS be_big NEG
'my house is not big' (ibe363-00.045)

(b) *àdòdó bó hu è → [àdòdó bóhujè]*
flower FUT dry NEG
'the flower will not dry' (ibe363-00.194)

(c) *néè de ɛnam è → [ńè dɛnamjè]*
flower FUT dry NEG
'the flower will not dry' (ibe363-00.036)

(d) *íntiǹ làg è → [íntiǹ làgìjè]*
nothing break NEG
'nothing broke' (ibe265-00.063)

As *è* is a clause-final particle, it usually is the last element of the clause. The only possible element to follow *è* is the emphatic particle *ò*. This combination is pronounced [jò] regardless of the preceding phoneme.

(3.198) *báà kɔn è ò → [baakɔni jò]*
3PL.HUM.NEG fight NEG EMPH
'they did not fight' ibe241-00.239)

3.7.6.1 Standard negation

This is the basic way to negate declarative verbal main clauses. Abesabesi's standard negation includes the clitic *è* and the lack of the 1H (Example 3.199). If subject pronouns are required, the irrealis pronouns must be used (Example 3.200, also Section 3.5.2.1).

(3.199) (a) *ebo H miʃ → [ebó miʃ]*
dog 1H sleep
'the dog sleeps' (ibe363-00.014)

(b) *ebo miʃ è → [ebo miʃìjè]*
dog sleep NEG
'the dog does not sleep' (ibe363-00.043)

(3.200) (a) *à H deŋ → [àdéŋ]*
1PL 1H climb
'we climbed' (ibe124-00.199)

(b) *asáà deŋ è → [asáà deŋjè]*
1PL climb NEG
'we did not climb' (ibe124-00.207)

3.7.6.2 Negation of other clause types

Besides declarative verbal main clauses, Abesabesi uses standard negation in a few "non-standard" clauses. Firstly, the only defunct verb that can be negated – *té* 'exist' – employs the standard negation described above. As the 1H never appears in affirmative use of defunct verbs (Section 3.4.4.3), the only difference between affirmative and negative sentences with *té* is the negation particle *è*.

(3.201) *iŋni té è*
water exist NEG
'there is no water' (ibe265-00.171)

Secondly, negated imperatives employ the standard negation. However, while affirmative imperatives drop the 2SG subject pronouns, their irrealis equivalents are obligatory in negative imperatives.

(3.202) (a) *ɲadum*
sit
'sit down!' (ibe070-00.073)

(b) *sáà ɲadum è*
2SG.NEG sit NEG
'don't sit down!' (ibe070-00.077)

Since predicate nominal clauses, attributive clauses, locative predicate clauses, and possessive clauses employ the verbs *tu* 'be', *wa* 'be (hypotetical)', *ʤe* 'be (names)', and *ʃi* 'have', they are negated like other verbal clauses using the standard negation.

(3.203) (a) *ɛsagum tu ehidino è*
sheep be black NEG
'the sheep is not black' (ibe365-00.181)

(b) *aséè ʃi ebo è*
1PL.NEG have dog NEG
'we don't have dogs' (ibe292-00.589)

Two clause-types, however, use additional morphological material in order to differentiate the affirmative from the negative: polar questions and conditional clauses. In polar questions, the particle *moní* is added before the subject NP (Section 3.7.5.1). *(mo)ní* has the equivalent *(mi)ní* in affirmative clauses but is used much more frequently than the equivalent.

(3.204) (a) *ɔ́ˡɔ́ ku?*
2SG.Q fall
'did you fall?' (ibe124-00.112)

(b) *moní séè ʤe iniŋ è?*
Q.NEG 2SG.NEG eat something NEG
'did you not eat?' (ibe265-00.250)

Conditional clauses add the low toned nasal *m̀* before the subject NP. As affirmative conditional clauses already use irrealis pronouns this adds further differences between affirmative and negative conditional clauses.

(3.205) (a) *m̀ eji jel ìʤɛ è*
COND.NEG rain fall today NEG
'if it does not rain today' (ibe265-00.287)

(b) *ǹnéè je u ìʤɛ è*
1SG.COND.NEG see 3SG.OBJ today NEG
'if I do not see her/him today' (ibe265-00.296)

3.7.6.3 Constituent negation

Single NP constituents of a clause can be negated through the negative focus construction *ɛ́ɛ̀ tu NP CLAUSE è* 'it is not NP who/which CLAUSE'. This construction is discussed in Section 3.8.2.

(3.206) *ɛ́ɛ̀ tu òsɛ na làg ikpánù ìdɛ́n è ò*
3SG.NHUM.NEG be father my break plate those NEG EMPH
'it is not my father who broke the plates' (ibe265-00.121)

3.7.6.4 Negative indefinite pronouns and polarity-sensitivity

Abesabesi exhibits some words or morphemes that either only occur within the scope of negation or express different meanings outside the scope of negation.

The non-specific indefinite pronouns (see Section 3.5.2.9 for the word class and Section 3.6.4 for their formation) are used in affirmative and negative contexts. In an affirmative scope, they are read as non-specific indefinite pronouns (e.g., *ɔ́ntɔn* 'whoever/anybody'), and in a negative scope, they are read as negative indefinite pronouns (e.g., *ɔ́ntɔn* 'nobody').

The irrealis pronouns are used in all negative contexts but cannot be counted as polarity-sensitive as they also occur in some affirmative contexts, such as with the affirmative habitual and in affirmative conditional clauses.

The perfect marker *ka/ke/ko* only occurs in an affirmative scope, while its negative equivalent *kinà/kinè/kinò* only occurs in negated clauses.

3.8 Information structure

This section describes several pragmatically marked constructions in Abesabesi that manage information structure. These are the focus construction, the negative focus construction, the relativized focus construction, the contrastive focus construction, and left dislocation. In order to indicate focused phrases in the translations of interlinear examples, they will be written in capital letters.

3.8.1 Focus construction

Abesabesi's focus construction contains a fronted NP, the focus marker *gí*, and the remaining clause with a gap for the fronted NP.[17]

(3.207) *ani na gí bɔ̀r atɔ*
people 1SG.POSS FOC hoe ground
'MY FAMILY hoed the ground' (ibe128-00.025)

[17] In order to indicate the focused phrase in the translation of the examples, it will be capitalized.

As Example 3.207 shows, the subject of the original clause, *ani na*, leaves a gap before the verb *bɔ̀r*. The gap is not filled with a resumptive pronoun. Focus fronted subject pronouns must be independent pronouns, as only these can head an NP (see Example 3.208).

(3.208) *ònì gí bɔ̀r atɔ*
1SG.IDP FOC hoe ground
'I hoed the ground' (ibe128-00.004)

NPs with the following syntactic functions can be focus fronted: subject (Example 3.209 a), primary object (Example 3.209 b), secondary object (Example 3.209 c), predicate nominal (Example 3.209 d), or adverbial (Example 3.209 e).

(3.209) (a) *kɔfí gí mə̀wɛ̀ íni kóko (ibe065-00.060)*
coffee FOC do_well and cocoyam
'COFFEE did well as well as cocoyam' ((ibe065-00.060)

(b) *ìtʃúgi gí ɔ̀ ʃɛ̀*
lie FOC 2SG tell
'you are telling LIES' (ibe105-00.138)

(c) *òkú iniŋ gí ɔ́ bá ki u*
all thing FOC 3SG.HUM FUT give 3SG.HUM.OBJ
'he would give him EVERYTHING' (ibe065-00.166)

(d) *áje no gí ɔ́ tu*
mother 1SG.POSS FOC 1SG be
'she is MY MOTHER' (ibe015-00.027)

(e) *iʃo ɔmis gí bò bo dʒo*
house king FOC 3PL.HUM PST give_birth
'I was born IN THE PALACE' (ibe065-00.020)

In affirmative predicate nominal clauses with the copula verb *tu* 'to be', the predicate nominal is frequently focus fronted (Example 3.210 a). Exceptions are content questions where the interrogative pronoun is usually situated at the beginning of the sentence, but without the focus particle *gí* (Example 3.210 b). Only rarely, *gí* is used after interrogative pronouns (Example 3.210 c).

(3.210) (a) *ohij u gí ǹ tu*
older_sibling 3SG.HUM.POSS FOC 1SG be
'I am her older brother' (ibe015-00.033)

(b) *máàn ɔ̀ tu*
who 2SG be
'who are you?' (ibe017-00.034)

(c) *màd gí tu ɛ̀dɛ́n*
what FOC be that
'what is that?' (ibe017-00.034)

Only NPs can be focus fronted. Yet, as NPs can be relativized, the focus fronted element can be quite heavy.

(3.211) *èdʒì mi bè mĩ ó jù u gí ǹ bá ʃa*
like REL they do PROG bury 3SG.HUM.OBJ FOC 1SG FUT speak
níkpa si ɛ èhen
about it now
'I will now speak about HOW THEY BURY HER/HIM' (ibe059-00.007)

Another method of focusing is the independent focus construction. This is a way to mark a single NP as the focus of a previous sentence or as the answer to a question without repeating the clause. The NP is simply followed by the defunct verb *gó* (see Section 3.4.4.3).

(3.212) *àdɔ̀m ísas gó*
month three FOC
'it is three months' (ibe167-00.165)

3.8.2 Negative focus construction

The focus construction presented earlier only applies to affirmative clauses. If the focus of a sentence is a negated constituent, the negative focus construction is used. This construction also is the only method to negate a specific constituent (Section 3.7.6.3).

This construction consists of the 3SG.NHUM.NEG pronoun *ɛ́ɛ̀*, the copula verb *tu* 'to be', the focused constituent, and the rest of the sentence ending with the negation particle *è*. The construction is thus similar to the English equivalent 'it is not..., who/which...' with the difference that Abesabesi does not employ a relative clause.

(3.213) *ɛ́ɛ̀ tu ònì ʃag atɔ è*
it.NEG be 1PL.IDP sweep ground NEG
'it is not me who swept the ground' (ibe128-00.038)

As discussed in the last section, affirmative equations and proper inclusions have a focus fronted predicate nominal by default[18]. Negative equations and proper inclusions, on the other hand, are rarely found in the negative focus construction (Example 3.214 a). The negative focus construction is thus not a default for negative equations and proper inclusions, but used to focus on the predicate nominal (Example 3.214 b).

(3.214) (a) *náà tu àgbɛ̀ è*
1SG.NEG be farmer NEG
'I am not a farmer' (ibe005-00.286)

(b) *ɛ́ɛ̀ tu olùkɔ́ ǹ tu è*
3SG.NHUM.NEG be teacher 1SG be NEG
'I am not A TEACHER' (ibe265-00.127)

[18]See Section 3.7.2 for a definition of equations and proper inclusions.

3.8.3 Relativized focus construction

A way to emphasize a verb and mark it as the focus of a clause is the relativized focus. It consists of a nominalized verb that is modified with a relative clause containing the rest of the clause. Its structure is discussed in detail in Section 3.6.1.2.

3.8.4 Contrastive focus construction

If the proposition of a clause is contrary to the presupposition of the interlocutor, Abesabesi marks it with the conjunction *ŋ̀gí*. This construction will be called "contrastive focus". Example 3.215 shows a situation where information is corrected by the speaker.

(3.215)

ɔ́ɔ̀	*á*	*kɔ*	*ìwé*	*è,*	*ŋ̀gí*	*ɔ́*	*á*	*miʃ*
3SG.HUM.NEG	PROG	write	letter	NEG	CFOC	3SG.HUM	PROG	sleep

'he is not writing a letter, he is sleeping' (ibe190-00.104)

3.8.5 Dislocation

A common way to initiate a clause with the topic is left dislocation. The dislocated phrase is separated by an intonational break from the rest of the clause. A resumptive pronoun replaces the phrase in the clause.

(3.216)

èdʒɛ́ʃ	*ekìn,*	*bà*	*ka*	*kèd*	*ɛ*
basket	one	3PL.HUM.SBJ	PRF	carry	3SG.NHUM.OBJ

'one basket, they took it' (ibe120-00.030)

Part II

Digital Grammaticography

Chapter 4

Grammaticography

Having set the base for the exemplary case study of this thesis in part one – the Abesabesi Sketch Grammar – I will now embark on pursuing the declared aim of this project – digitizing the format "grammar". While the following chapter will tackle the digitization itself, this chapter will comprise an in-depth investigation of the object of research – the grammar. It aims at the isolation and analysis of the format "grammar" in order to model it later in Chapter 6. Only if the format itself is thoroughly examined can it be clearly defined and distinguished from other formats. This will facilitate the isolation of a common structure that can serve as a model for the digital format.

This chapter therefore deals with "grammaticography", the art and practice of grammar writing. This discipline has a long tradition, reaching all the way back to the earliest known grammar: Panini's grammar of Sanskrit (von Böhtlingk, 1964). To differentiate the activity of grammar writing from the study of grammar writing, Lehmann and Maslova (2004, p. 1857) coined the term "metagrammaticography", which is the field this chapter can be attributed to. The study of how to write grammars has possibly preoccupied grammar writers since the date that the first grammar has been written. Only recently though, it has emerged as a linguistic subfield within the language documentation realm, where great numbers of new grammars are produced. An extensive amount of papers have been written within this field, some of them compiled in one of the books solely dedicated to grammaticography (Zaefferer, 1998; Ameka, Dench, & Evans, 2006; Payne & Weber, 2007; Nakayama & Rice, 2014).

Along the lines of the existing research in metagrammaticography, this chapter will collect definitions, typologies, and structures of grammars following this outline:

1. Define "grammar" (Section 4.1)
2. Categorize grammars and define general parameters (Section 4.2)
3. Choose the kinds of grammar to be digitized (Section 4.2)
4. Examine the structure of grammars (Section 4.3)
5. Describe the grammar in context (Section 4.4)

6. Investigate digital grammaticography (Section 4.5)

7. Collect and analyze previous approaches to digital grammaticography (Section 4.6)

In order to isolate and analyze the format "grammar", Section 4.1 defines the concept "grammar" and establishes a notion that will be used throughout this thesis. Different types of grammars will be introduced in Section 4.2. These types will be categorized according to general parameters before selecting the parameters that will confine the object of digitization. Knowing the type of grammar relevant to this project will be essential to model the structure of grammars (Section 4.3). While diverging structures will be traced back to different needs and values, common structures can be used to model a digital format. In addition to the internal structure of grammar, its relationship to other resources (dictionaries and text collections) needs to be examined in order to find possible points of interaction that a digital format needs to provide. These external resources will be discussed in Section 4.4. Later in this chapter, grammaticography will be transferred to the digital sphere. Section 4.5 investigates the field of digital grammaticography and defines the notion "digital reference grammar (DRG)". Previous approaches to the digitization of grammars will be listed and evaluated in Section 4.6.

With the majority of languages being undescribed or underdescribed (Payne, 2005b, p. 236) and plenty of languages being endangered (Krauss, 1992; Crystal, 2000), there is an urgent need for descriptions of more languages as part of their documentation. For these reasons, the setting for the digitization project will be within the language documentation paradigm and concentrate on grammar writing for un- and underdescribed languages. Moreover, the scope of this project will be confined to a type of grammar that is generally referred to as "reference grammar" – a descriptive grammar targeted at an academic audience of linguists with an interest in typology, the language family, or the linguistic area. This scope and its parameters will be further discussed in Section 4.2.

4.1 What is grammar?

Let us examine the term "grammar" from different angles. I would like to highlight three different concepts encompassed by this term, which are each objects of research for a separate branch of linguistics.

If asked about "grammar", a computational linguist would probably give you a definition along the lines of the following statement – inspired by Becerra-Bonache, Bel-Enguix, Jiménez-López, and Martín-Vide (2015, Section 9.2.3).

Concept 1 (Grammar) *A grammar is a set of production rules over an alphabet that form all valid sentences of one language.*

These formal grammars are used to model correct syntax in all kinds of documents and especially in programming languages where the correctness of the syntax must be checked before code can be compiled and executed. Applied to natural language – famously

done so by Chomsky (1957) to support his linguistic theories – they find use in linguistic technologies that require syntactic parsing such as machine translation. One should bear in mind that this definition sees grammar as a set of rules that are applied to a lexicon (in formal languages, it is called an "alphabet"). A language only functions through the interaction between these two distinct sets. No sentences can be generated as long as we lack a lexicon. The relation between grammar and lexicon will further be discussed in Section 4.4. Apart from these thoughts, however, we will exclude this concept of grammar from further discussion.[1]

The following concept is the subject of multiple syntactic theories, such as Radical Construction Grammar (Croft, 2001), Cognitive Grammar (Langacker, 2008), and Optimality Theory (Kager, 1999), where hypotheses about the internal structure of grammatical information are proposed.

Concept 2 (Grammar) *Grammar is the set of rules or constraints that enable speakers to combine words to make well-formed sentences.*

The internal structure of grammar is an important concept for this undertaking as it could be the template for how authors should structure a grammatical description. Linguistics theories, however, differ greatly in their perspective on how grammatical knowledge is stored and processed by the brain. Moreover, the field of psycholinguistics finds there is a lack of neurolinguistic evidence to answer the question of how this knowledge is stored and processed (Rizzolatti & Craighero, 2007, Section 47.4). Concept 2 will thus not be considered in further discussion.

The third concept relates to the above Concept 2 in that "describing" a language tends to presuppose that this internal set of rules is what is being described. This third concept would probably first come to mind for those interested in typology or descriptive linguistics.

Concept 3 (Grammar) *A grammar is a book that contains a description of a language's structure.*

Concept 3 is the concept relevant to this attempt to digitize the format of grammars. However, we ought to apply a major change in order to have a definition that suits our project. As this thesis intends to challenge the conception of grammatical description being tied to the format of a book, the word "book" needs to be replaced by a more general term: "medium". Moreover, grammatical descriptions are never able to cover an entire language and, especially when working with minoritized languages, a standard variety to represent the whole language does not exist. Therefore, the term "language" needs to be replaced by "language variety". Having chosen the concept of grammar relevant to this project and having refined the wording, a first definition has been chosen to be used for the attempt of modeling grammars.

[1] Consult (Maxwell, 2011, 2012) for the relevance of computational grammars for digital grammaticography.

Definition 1 (Grammar) *A grammar is a medium that contains a description of a language variety's structure.*

This definition also distinguishes the terms "grammar" and "(grammatical) description", as they will be used henceforth. While the term "grammar" is used to refer to the medium itself or the literary genre, the term "grammatical description" signifies the content of a grammar.

In the following section, the term "grammar" as a literary genre will now be confined to a specific subgenre in order to narrow down the scope of this digitization project. Parameters of grammar writing will be introduced in order to compile a typology of grammars and then to chose the subgenre of interest.

4.2 A typology of grammars

Grammar writing depends on various parameters, causing different grammars to be quite diverse in form and function. Much has been written in terms of how grammars differ, so this section summarizes the key parameters presented in literature.

The following parameters combine and extend those of Rice (2005) and Mosel (2006):

- Audience
  - Linguists
    - Experts in the language family or linguistic area
    - Experts in a linguistic theory
    - Typologists
  - Language speakers
  - Learners/Teachers
  - Non-linguist researchers
- Purpose
  - Impart linguistic knowledge
    - Typology
    - To further formalistic theories
    - Comparison within language family or linguistic area
  - Language learning
- Speaker community
  - Wishes and needs
  - Number of available consultants

- Available resources
  - Number of researchers
  - Amount of time
  - Amount of money
  - Amount and type of fieldwork
  - Amount and type of previous work/documentation
- Extent
  - Full reference grammar
  - Short grammar
  - Grammar sketch
- Metalanguage
  - Monolingual
  - Bilingual
  - Multilingual
- Description vs. Prescription
  - Descriptive
  - Prescriptive
- Universalism vs. Particularism
- Variation
  - Description of a standard
  - Description of one or more varieties
- Language status
  - Description of an endangered language
  - Description of a vital language

Audience and Purpose These two parameters are interconnected as the purpose usually defines the audience. If a grammar is written to teach a language, the main audience will be teachers or learners. Grammars for un- and underdescribed languages are mainly written for academic purposes and therefore attract an academic audience of linguists (Noonan, 2005, p. 353). These grammars, however, can be remodeled for teaching purposes (Mosel, 2006, p. 42). Of the three linguistic purposes, the most controversial might be to further linguistic theory. The role of language theory in grammar writing has in

fact been heavily discussed in the literature (e.g., Leitner & Graustein, 1989; Rice, 2005). As the purpose of promoting formalistic theories will be excluded from the scope of this project, this topic will only be of marginal interest. However, theory in grammar writing will be of concern for the content of grammars (Section 4.3.1) and the use of terminology (Section 6.3.3.2). Another point that is connected to the audience as well as the purpose is whether the grammar is designed as a database or as a coherent story. These two extremes open up a scale that correlates with readers that like to pick desired information out of the grammar on the one hand and those that like to understand the grammatical structure of a language as a whole on the other hand.

Speaker community The speaker community is not only an important actor in a language documentation project but also in the process of grammar writing. As grammar writing is dependent on the data provided by the speaker community, factors such as the number of consultants, the degree of consultant participation in the project, and the documented speech genres can shape quantity and quality of a grammar. Ethical fieldwork and research presupposes respecting the wishes and needs of the community. This includes, for example, how the community is represented in the documentation corpus or even through examples in a grammar. Speakers' participation in grammar writing can be facilitated through linguistic training or with tools such as Summer Institute of Linguistics (SIL)'s PAW (C. A. Black & Black, 2012). Grammars can also be produced entirely by native speakers. Writing a grammar about one's native language can entail advantages and disadvantages, as well as writing a grammar as a non-native speaker (Ameka, 2006). This digitization project will be aimed at grammar writers that are trained linguists. Therefore, it will not attempt to speak to the needs of grammar writing for native speakers without linguistic knowledge (as the PAW tool does). Even though this thesis is aimed at trained linguists, the participation of the speaker community should not be limited due to the structure of the DRG format.

Time and Extent Again, these two parameters are interconnected. The more time you have at your disposal, the bigger the extent of your grammar can be. The extent of a grammar can be seen from a qualitative or a quantitative perspective. More time and resources can either lead to a more comprehensive grammar covering many topics, or to a grammar with only a few topics investigated in depth. On the one hand, constraints on time decide on how much fieldwork can be undertaken and how long a researcher can spend on the analysis and writing of the grammar. On the other hand, time other people have spent on researching the language – in the form of language documentation or description – determines how a grammar will be shaped. The extent of a grammar project should therefore be planned beforehand, taking into account the amount of fieldwork that is possible and the amount of existing work on the language. The scale stretches from a grammar sketch that only contains basic typological information and the most important phenomena of a language to a full reference grammar that "grows" over years in the sense of Weber (2005b). Grammars of un- or underdescribed languages, which are the target

of this digitization project, can also differ in time and extent. They can be written in the scope of a smaller project, resulting in a grammar sketch, such as the Abesabesi Sketch Grammar used as a case study in this project, or in the scope of a life-long project that eventually is published. Therefore, our digitization project should encompass different types of length and be applicable to sketch grammars, as well as long reference grammars.

Metalanguage Grammatical descriptions can contain multiple languages. A distinction which will be important for the grammar format is the distinction between the language that is described and the language that is used to describe the other. The described language will henceforth be referred to as the "target language" and the language used to describe it will be referred to as the "metalanguage". Choosing a metalanguage for the grammatical description depends on the audience. As we have chosen an academic audience of linguists, we expect the metalanguage to be a language of wider communication. English, for example, is often chosen as a language of global communication, especially in academia. However, grammars are not always written in English but sometimes in another language of wider communication, in the local lingua franca, or perhaps even in the target language (monolingual grammars). This digitization project will be open to any metalanguage – and also be open to the possibility that the grammar may be (re)written in multiple languages.

Description vs. Prescription As we focus on a linguistic audience, we will only consider descriptive grammars. Prescriptive grammars might be useful as learners' grammars or for language standardization purposes. These genres, however, will not be considered in this project.

Universalism vs. Particularism As Lehmann and Maslova (2004, p. 1859) point out, grammar writing happens on a scale between universalism and particularism. Some grammar writers follow the goal to stress a language's uniqueness. They produce grammars with rather ideosyncratic structures that are harder to understand. Other writers follow a universalist perspective and describe the language in terms known from other languages. These descriptions tend to be less interesting, but easier to compare. This scale is a key problem for every attempt at standardizing grammars. As standardizations generally suppresses particularism, means must be provided for a grammar author to stress a language's uniqueness.

Variation and status Considering the amount of variation and the degree of standardization becomes very important when working with un- or underdescribed languages. As these languages are often not spoken by peoples with economic power and, therefore, rarely fill administrative domains, a standardized variety probably does not exist. Grammar writers thus must decide on whether they want to only describe one variety or the whole language, mentioning linguistic variation when it occurs. The former is obviously more time-consuming if not impossible and might affect the descriptive accuracy when

variation is not clearly marked, as Rice (2005, p. 397) points out. It is thus desirable for our project to offer some tool to differentiate linguistic varieties and a way for readers to filter the descriptions for varieties. Moribund languages with only a handful of speakers left might only allow for the grammatical description based on one ideolect. This parameter is thus also dependent on the type of fieldwork and the consultants available.

Summing up, I will consider the following parameters in this undertaking:

The format to be developed will consider grammars that:

- are produced within a **language documentation project**
- are for an **academic audience of linguists**
- have a **linguistic purpose** (such as typology, research in a language family, etc.)
- cater to the needs and wishes of the **speaker community**
- are for researchers with **different amounts of resources** available
- have **any extent** (from a grammar sketch to a full reference grammar)
- are **mono-, bi-, or multilingual** and use **any metalanguage**
- are **descriptive** and allow for ways to **stress the language's uniqueness**
- describe **any amount of varieties** with **any status** (moribund, endangered, vital, etc.).

The parameters that have been chosen lead us to a subgenre of grammar, which is often referred to as "reference grammar", "descriptive grammar", or "linguistic grammar" in literature. The term "grammar" will be be used from here on to denote grammars adhering to the presented criteria.

4.3 The structure of grammars

Grammars can differ to a large extent. Some are based on "universal" categories, others try to stress the language's "uniqueness" by using an idiosyncratic structure. Some have a heavy focus on "contemporary" topics, others attempt to cover all "traditional" topics in linguistics. In order to model the format "grammar", the structure of grammars has to be analyzed and dismantled. A first step is to extract the individual content, the logic of the grammar, and the order of topics from the formal structure. While the formal structure of a grammar defines common components that can be assembled in a flexible manner, the content, logic, and order are largely dependent on the individual grammar. The formal structure will be modeled in Section 6.3, before proposing a digital format. Content, logic, and order, on the other hand, are only marginally important to modeling grammars as they depend on the ideas of an author or sometimes the guidelines of a

publisher. They are, however, important to consider, as they constitute the individual essence of a grammar and transfer a language's "genius" in the sense of Sapir (1921, p. 127). Therefore, a digital format should not impede the author in their creativity by presetting the order, logic, or content of a grammar. Section 4.3.1 will deal with the content of a grammar and the outline templates proposed by several authors. The order of grammar entries will be analyzed in Section 4.3.2 based on three different parameters: complexity, regularity and cross-referential dependence. Finally, Subsection 4.3.3 discusses the logic of a grammar by presenting different approaches to grammar writing in terms of the form-function relation represented by a grammar entry.

At first, however, it is important to introduce the notion of a "grammar entry", which is a central unit of the formal structure. It will be discussed in detail in Section 6.3.1.3), but as it is also vitally important to this discussion, a brief introduction is given here. The grammar entry subsumes the terms "chapter", "section", and "subsection". It contains the description of a single phenomenon, which can then be divided into sub-phenomena described in other grammar entries. Entries are thus the central unit in grammar writing by forming a complex hierarchical network that make up the grammatical description.

4.3.1 The content of a grammar

Recommendations for the ideal content of a grammar have been given from the beginning of grammar writing. However, no recommended outline has succeeded to define the standard structure of grammars. This can be traced back to the fact that grammar outline templates are influenced by current trends in linguistic research. The more fine-grained an outline is, the more controversial and the less sustainable it becomes.

Hence, using outline templates as durable constraints for a digital format of grammar does not seem to be recommendable. Furthermore, outline templates are often criticized for their rigid structure into which languages are being pressed. This stands in stark contrast to Rice's (2006) recommendation to let a language tell its own story. She propagates a structure that is centered around the language's unique structure rather than along the lines of an outline template. Weber (2005b, p. 430) also recommends the use of outlines with caution. Nevertheless, templates must be considered in this project as they have always guided grammar writers through the process of structuring and writing a grammar. Outline recommendations can be given from two different perspectives: defining the minimal set of topics to be included in a grammar or defining the maximum set of topics by providing an outline containing all possible topics.

A minimal list of essential topics is recommended by Noonan (2005, p. 360) and Rice (2005, p. 395), who both state that a good grammar contains at least the traditional topics: phonetics, phonology, morphology and syntax. Pawley (2014, p. 14) adds semantics to the list. Another way is to provide lists of more detailed topics not grouped into the traditional linguistic levels (e.g., Mosel, 2006; Pawley, 2014). These "minimal outline templates" can be a great tool for grammar writers to check whether an essential topic is missing in their grammar. However, they differ greatly, which urges the grammar writers

to choose their own preferred minimal outline. With regard to the digitization of grammars, we could provide an array of minimal outlines to choose from that would function as a writer guideline. However, using them as a constraint for the minimal extent of a grammar would simply be impossible, as they differ to much.

Maximal outline templates, on the other hand, usually contain a complex chapter structure with (sub-)sections dedicated to more detailed topics of language structure. Moreover, they provide an ordering of the chapters and, thus, recommend a structure. A famous outline template is the Lingua Descriptive Questionnaire (Comrie & Smith, 1977), which has been the outline for a whole series of grammars. However, it has gained quite a bit of criticism for its contemporary nature, for the lack of certain topics, and for its overall chapter structure (e.g., Mosel, 2006, p. 57). Other example outlines are Mosel (1987) or Payne (2014, Appendix). Like minimal outline templates, maximal outline templates can guide the author to detect topics they might have not touched yet. Since maximal outline templates attempt to cover all topics necessary to understand a language's grammatical structure, they could be used to measure a value that (Good, 2012, p. 13) refers to as "completeness". A DRG that guides authors in the writing process could use maximal outline templates to indicate whether a grammatical description is "complete". However, completeness can only be measured as completeness at the time of publication. This is because new linguistic topics may arise in the future and become important for grammatical description. In light of the idea of growing a grammar (Weber, 2005b) and the possibility of distributed authorship (Good, 2012), the value of completeness becomes more important. This value is discussed in more detail in Section 5.1.6.

Introductory chapters of a grammar are usually located at the beginning of a grammar. They do not contain a description of the language's structure but give context about the language, its speakers, and the grammar itself. Mosel (2006, 46) mentions the following topics to be included in the introductory chapters: information about language, genetic relationships, socio-linguistic background, varieties described, theoretical approach, sources of data, methods of text analysis, and the conditions surrounding how fieldwork was carried out. From what Noonan (2005) recommends to be part of a grammar, a typological sketch and information on previous scholarship could be added to the introductory chapters.

4.3.2 The order of a grammar

The order of grammar entries essentially shapes a grammar, as it marks the course of a grammar's "story". Generally, the order can be based on the following criteria: complexity, regularity, and cross-referential dependence.

4.3.2.1 Complexity

There are two opposed approaches in structuring a grammar based on the complexity of the linguistic unit described: top-down (descending) vs. bottom up (ascending). Most

grammars are written in the bottom-up approach, which probably originated within the Latin grammar tradition (Mosel, 2006, 48). Starting with the smallest linguistic units (phonemes), the description adds complexity until the sentence level is reached, resulting in the following structure: phonology > morphology > syntax. The top-down approach follows the opposite direction, breaking the biggest unit into smaller pieces, until the smallest unit is reached. This approach has been followed rather strictly by the Lingua Descriptive Grammar Series. Scholars such as Rice (2005, p. 396), Mosel (2006, p. 57), and Payne (2014, 94) recommend a bottom-up approach.

4.3.2.2 Regularity

Likewise, most scholars plead for the description of regular patterns before irregular patterns, as irregular patterns might confuse readers without the knowledge of respective regularities (Payne, 2014, p. 94). A grammar reader will be more likely to understand a phenomenon if it is presented by introducing a regular pattern before adding all irregularities that human languages naturally bear. Both parameters, complexity and regularity, can be applied to the internal structure of a grammar entry as well as to the relative order of the entries.

4.3.2.3 Cross-referential dependence

A third parameter to bear in mind is the amount of cross-references necessary for the description of a phenomenon. As most of the phenomena in languages are hardly describable without referring to other grammar entries, a good starting point for each grammar chapter are those entries requiring a minimum of cross-references (Mosel, 2006, p. 43). This should also be reflected in the order of entries: “Basically, if an analytic decision concerning category X needs to refer to some facts concerning category Y, then the chapter dealing with Y should be positioned before that dealing with X” (Rice, 2005, p. 401).

Each of these parameters opens up a scale. Those grammars following extreme approaches, such as the Lingua Descriptive Grammar Series, tend to force a language into a radical order rather than writing a grammar inspired by the language’s structure. Digital grammars should therefore provide template structures that guide the grammar writer but retain the flexibility to change a template according to the structure of a particular language.

4.3.3 The logic of a grammar

Like any linguistic sign, grammatical signs also have a form and a function. A grammatical description, at least the morphosyntactic part of it, thus describes the relation between a form and a function[2]. As multiple forms can be related to one function and multiple

[2]Mosel (2006) provides a historical overview of the distinction between form and function in grammar.

functions can be related to one form, the question arises whether a grammar should be structured by forms or by functions. One way would be to list the existing forms and to describe each of their functions (semasiologic approach). The other one would be to list functions and to describe which forms are used to fulfill each of the functions (onomasiologic approach). The choice between a semasiologic and an onomasiologic approach is what I refer to as the "logic of a grammar".

4.3.3.1 Semasiologic approach

The overwhelming majority of grammars are structured semasiologically (Cristofaro, 2006, p. 141). This approach is perceived as a much more evident way of describing a language because form is what can be grasped directly. Moreover, Payne (2014, p. 101) argues that it is easier for the author to identify structures than functional categories. On the other hand, a form-based structure, as much as it highlights the uniqueness of a language, is harder to use for comparative reasons. Furthermore, it splits functional systems, as one function might be represented in different word classes or structural levels (Payne, 2014, p. 100). The TAM system in particular might span over different structural levels. A strict semasiologic approach would split the tenses and describe them in different parts of a grammar.

4.3.3.2 Onomasiologic approach

This approach is associated with linguistic typology, as it defines functional categories and describes which structures are employed in a particular language to perform a certain function. A functional grammar's starting point is thus a cognitive or communicative domain and its goal is a structure within the language. Consequently, similarities in how grammars express a certain meaning are more easily comparable. In contrast to a semasiologically ordered grammar, functional grammars rely on a universal ontology of functional concepts, as they are not necessarily given by the language. Such an ontology of functional concepts can be used as an outline template for onomasiologic grammars. However, only a few of these ontologies have been proposed (e.g., Comrie, Croft, Lehmann, & Zaefferer, 1993). Likewise, purely onomasiologically structured grammars are very rare. One example is *A communicative grammar of English* by Leech and Svartvik (1975).

Payne (2014, p. 101) stresses that a functional approach brings together "functional systems", linguistic strategies from different structural levels, which have a common purpose in language. On the other hand, the onomasiologic approach splits equal or similar forms with different functions although they might have a polysemic relation (Mosel, 2006; Cristofaro, 2006).

4.3.3.3 Pure approach

The separation of form and function entails a few problems, which is why grammar writers rarely use one of the mentioned approaches in a pure fashion. Lehmann (2004a, Section 3) mentions these three problems:

First, a fine-grained hierarchy of functional categories can hardly avoid categories that are defined through formal parameters. In the same way, in a semasiologically structured grammar, some formal categories must include functional parameters in order to be distinguished from other categories.

Second, a semasiologic description treats a form and mentions all functions it expresses. Any given function might also be mentioned in the entry about another form, as functions can be expressed through multiple forms. The question arises where to give a more detailed introduction about the respective function. This problem can likewise be applied to the onomasiologic approach.

Third, there might be restrictions or conditions in a language on when to associate a certain form with a specific function. These restrictions apply to the form-meaning pair and not to one function or one form. Where is a grammar writer thus supposed to describe these restrictions and conditions?

4.3.3.4 Combined approach

Most scholars, however, advocate a combined approach, where these problems do not occur. Three options are possible:

First, the author could follow their preferred logic of grammar approach but include an introduction to each chapter where they give a brief account of the section contents but from the perspective of the other approach. An onomasiologically structured grammar could, for example, have an introductory section for every chapter where all forms occurring in that chapter would be introduced (cf. Lehmann, 2004a, senction 3). Second, certain domains could be tied to one approach, whereas others could be tied to the other approach. This has been suggested rather abstractly by Noonan (2005, p. 358): “In general, it’s best to describe morphology with a form-to-function orientation and syntax with a function-to-form orientation”. Other scholars, such as Payne (2005a, p. 360), provide detailed lists of domains that go with a particular approach. Mosel (2006, p. 60) advocates a semasiologic approach for the main part of the grammar and an onomasiologic approach for additional chapters treating functionally oriented topics like “possession, orientation in space and time, and negation, and the pragmatic function of questioning”. Third, a grammar could contain a full semasiologic and a full onomasiologic account. As Cristofaro (2006, p. 147) points out, this might result in redundancies. Nordhoff (2008, p. 304), on the other hand, argues that real redundancies do not occur as both approaches treat the same form-meaning relationship but describe different aspects of it. The only remaining redundancies would be the examples. His grammar of Upcountry Sri Lanka Malay (Nordhoff, 2009) illustrates this. The grammar structure proposed

in Nordhoff (2012b)[3] draws on the construction grammar approach proposed earlier by Lehmann (2004a). Form-meaning pairs (FOMPs) would be the focus of a description. Separate entries exist for the form and the function but both entries reference the respective other. Through these references, a network of FOMP relations would account for poly-functional forms and functions expressed by multiple forms. This network could be used to structure a navigation tool such as a modified index that could visualize the links between form and function (cf. Cristofaro, 2006, p. 147).

4.4 Grammars in context

Having analyzed the internal structure of a grammar in the preceding section, this section now examines grammars in relation to other resources such as dictionaries and text collections. It is important to replicate the way in which paper grammars and these external resources interact in a digital format. In fact, a DRG will not only replicate but enhance these links in that readers will be able to quickly access the referenced resources via hyperlinks as opposed to browsing through yet another book. A grammatical description is impossible to produce without a collection of texts in which the described phenomena occur and a dictionary where the meaning of lexemes are listed. This set of grammar, dictionary, and text collection is often referred to as the "Boasian Triade", following the structure of the earliest fieldwork within the academic fields of ethnography and linguistics by Franz Boas. In fact, these three components of a fieldwork setting are heavily inter-dependent as explained in the following subsections.

4.4.1 Dictionary

The borderline between a dictionary and grammar is less clear than it seems at first. In fact, the border does not exist as a fact of the language itself but is a construct of linguists and the nature of linguistic study (Mosel, 2006, p. 46). Therefore, it is often a difficult decision whether to assign a certain phenomenon to the dictionary or to the grammar. Although we could follow the rule of thumb and mention function words in a grammar and list content words in the dictionary, this distinction is not obvious (cf. Schultze-Berndt, 2006). Other problems arise with phrasal or idiomatic expressions, which may also convey grammatical meaning (Kay & Fillmore, 1999). Phrasal and idiomatic expressions seem to be avoided by dictionary and grammar writers as they do not seem to fit in either of the categories (Mosel, 2006, p. 46; Pawley, 2014, p. 17) and would rather be expected in a learner's textbook. Mosel (2006, p. 46) relates this distinction problem to the fact that grammaticalization and lexicalization are continua.

As this kind of problem exists, it is all the more important to establish some kind of interaction between grammar and dictionary. An interlinked approach of both works is suggested – amongst others – by Weber (2005b, p.423) and Drude (2012). But even if

[3]This model is discussed in detail in Section 6.1.3

there is not much interaction, consistency in terminology is important across dictionary and grammar, as both works will probably be read by the same readers (Weber, 2005b, p.423).

4.4.2 Text collection

Collecting texts, preferably in different genres, precedes the process of grammar writing. No grammar can be written without having analyzed the structures within utterances. When publishing the grammar, the readers benefit from an accessible collection of texts where they can examine the described structures in context and might even find structures that have not been described (Noonan, 2005, p. 359; Mosel, 2006, p. 53). As grammar writing is influenced by contemporary findings of linguistic theory, the grammar might lack the description of structures that were not en-vogue at the time of writing (Mosel, 2006, p. 53). These structures can still be identified in a text collection. In paper grammars, a selection of texts is sometimes included as an appendix. This is, for example, an integral part of the Mouton Grammar Library series. Other authors have decided to publish a whole text edition along with the grammar (e.g. Mosel, 1977; Heath, 1980).

Digital grammars could benefit from a connection to corpora stored in language archives. These corpora are common products of language documentation projects and ideally contain transcribed, translated, and annotated texts along with their audio and video recordings. Furthermore, examples used in the grammar could be linked to the recording if they stem from actual utterances of the corpus.

4.5 Digital grammaticography

As mentioned earlier, this chapter deals with grammaticography – the art and practice of grammar writing – which is itself situated in the field of metagrammaticography – the study of grammar writing. This section will now explore a particular form of grammaticography, "digital grammaticography"[4], which is the art and practice of grammar writing using a digital medium. This research is therefore an advance into "digital metagrammaticography" – the study of grammar writing via a digital medium. This section examines the product of digital grammaticography – a digital grammar. As the scope of this project is confined to the subgenre of reference grammars, the object of research will be the "digital reference grammar (DRG)"[5].

A grammar involves three processes: grammar writing, storing the grammar, and reading the grammar. Metagrammaticography is mainly concerned with grammar writing, as it is supposed to discuss questions such as "How can grammars be structured?"

[4] An earlier term for this concept is "electronic grammaticography" (e.g., Nordhoff, 2012a).

[5] This concept has earlier been called "electronic reference grammar" (Zaefferer, 2006; Nordhoff, 2008). Another term used in the field of digital metagrammaticography is "hyptertext grammar" (e.g., Bender, Ghodke, Baldwin, & Dridan, 2012). This notion also specifies the medium of the grammar (hypertext pages).

and “What needs to be in a Grammar?”. At least indirectly, but often also directly, it deals with the process of reading grammars, as it is concerned about the audience and about how to transmit information to the reader in a comprehensible manner. The second process, storing the grammar, however, is completely ignored in metagrammaticography. The reason for this might lie in the simplicity of storing an analog grammar: it is printed on paper, bound to a book, and stored in libraries or on private book shelves. As digital grammars rely on a different medium, the process of storing the grammar is also completely different. Digital metagrammaticography needs to address this topic and propose methods for storing digital grammars. The other two processes, writing a grammar and reading the grammar, differ when a digital medium is used. The author requires a tool to write the grammar and the reader requires a tool to open the grammar.

The process of storing the grammar will be the main focus of this thesis, as it influences the other two processes. The way the grammar is stored influences the way the grammar can be written or read. In order to store the grammar, a document or a collection of digital documents is required. The digital documents need to be stored on a device, where it is optimally accessible from all over the world. The best option for easy global distribution to date is probably the internet. The digital documents need to adhere to a standardized and machine-readable format in order for present and future software to modify or read them. A digital format for grammars will be proposed in Chapter 6.

In order to write or read the grammar, tools will be required that interact with the digital document and display or modify it. The first required tool assists writers to write or modify the grammar. It guides the author to write in a format-conforming way. Any tool with these functionalities will henceforth be called a “grammar writing tool (GWT)”. The primary purpose of the proposed format is the encoding of “digitally born” grammatical descriptions. Therefore, a GWT will be regarded as an integral part of a DRG. This project is only marginally concerned with “retrofitting” paper grammars – that is, digitizing existing paper grammars. That means, although the format is designed for digitally born grammatical descriptions, it can also be used to re-encode existing grammars.

Grammar readers require a second tool that reads the document and displays it in a reader-friendly layout with navigation tools and search functionalities. Any tool with these functionalities will henceforth be called a “grammar exploration tool (GET)”. A GWT and GET could also be combined in one application with separated access for writers and readers. For the theoretic purposes of this thesis, however, they will be treated as two separate tools. A combined approach will, however, be presented in Section 5.4.4.

In summary, I regard a DRG as the combination of three interacting parts: the grammar document in a digital grammar format, the GWT, and the GET. While the grammar document contains the grammatical description, the GWT is an application that can be used to fill the grammar document with content or to modify it. The GET displays the content of the grammatical description in a readable way. The basic structure of a DRG can be seen in Figure 4.1

This PhD project compasses the development of a GET. This will exemplify how an

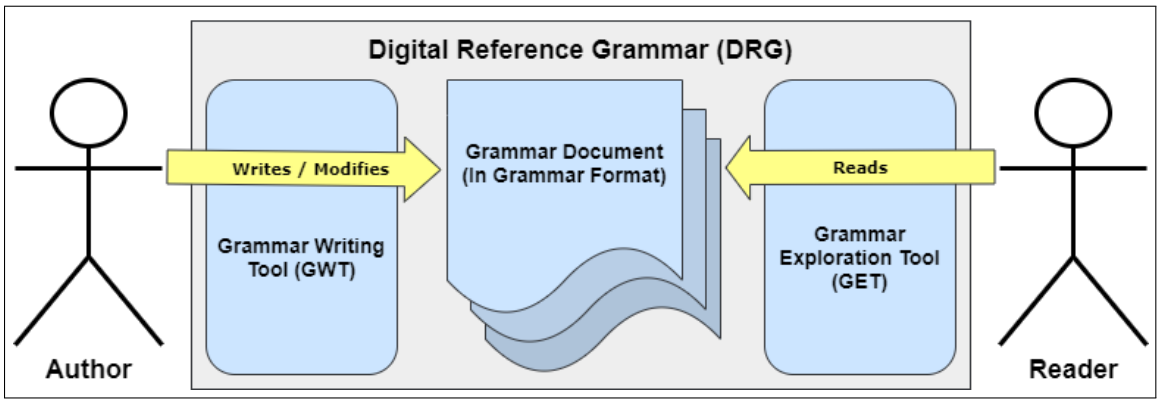


Figure 4.1: The basic structure of a DRG

application interacts with the digital format and will also test the limits of the digital format (Chapter **??**). However, the DRG will only feature the functionalities that could be developed within the scope and time frame of this project. Nevertheless, further functionalities will be considered in the format's evaluation and discussed in Section 7.4. A GWT will not be developed. It will, however, theoretically be discussed in this chapter. I believe that the development of both a GWT and a GET is a task that is only achievable with the support of a larger institution such as a digital publisher or a language archive. This will be the topic of Chapter 8.

Having introduced the object of research – a DRG – and outlined its basic structure, previous approaches of conceptualizing or developing a DRG will be presented in the following section.

4.6 Previous approaches to digital grammaticography

Both in theory and in practice, approaches to digitize grammars or to produce DRGs have lead to several academic papers and a few actual DRGs. This section presents these approaches and categorizes them. Many approaches have only tackled parts of what was defined as a DRG in the previous section. Those approaches which only define a format for the grammar document will not be presented here but later in Section 6.1 before proposing my own format for DRGs.

In fact, there are rather simple methods of digitizing grammars or producing them digitally using existing technologies. For example, a paper grammar could be scanned. The result of this approach does not differ too much from paper grammars, as it barely has any additional functionalities. Furthermore, digital text documents such as grammars that have been typed on a computer using a text-editor could be counted as a DRG. While the text editor is both a GWT and a GET, the text file is the grammar document. Digital

text formats can easily be distributed through the internet and have the advantage of a text that is is machine-readable and therefore facilitates full text search. Additional functionalities of this approach, however, are also limited.

4.6.1 PDF approach

Especially popular for digital publishing is the open standard Portable Document Format (PDF), which can be read and modified by many proprietary and open-source applications. Many published grammars can be purchased or accessed online by downloading a PDF document. As PDF has become an open format, it does not depend on a specific application. Therefore, the ease of creating or editing a PDF document heavily depends on the editor used. Texts are usually edited in a text editor before the editor's internal format is converted into PDF for publishing. For example, Microsoft's widely used text editor Word has a rather flat learning curve, as it lets users edit texts while displaying the appearance it will have as a finished product. These kinds of editors are generally called "What You See Is What You Get (WYSIWYG)" editors. Microsoft Word saves the text in its own format – the Office Open XML Document format – with the file extension `.docx`. Before publishing, the editor can be used to convert the document into the PDF format. Using a semi-logical mark-up language like (La)TeX (Knuth, 1984), which can be edited in different editors, requires a longer learning process but provides more functionalities and allows for more control over the formatting process.

The advantages of the PDF format are that it is an open standard and is therefore suitable for archiving (value "longevity" in Section 5.1.7). The use of hyperlinks allow for a fast navigation within the document and to sources outside the document (value "ease of information retrieval" in Section 5.1.1). Audio and video files can also be embedded in a PDF document and played using an editor or browser. This way, every example can be illustrated with an actual utterance (value "accountability" in Section 5.2.1). However, the longevity of embedded audio and video files cannot be secured like the longevity of the PDF format itself. A good example is the withdrawal of support for Adobe's Flash-player, which means that old PDF files that have embedded Flash videos cannot be played anymore.

PDF is a format for presenting documents independent of hardware, operating system, or software. That makes it suitable for publishing and printing but less for modifying or extracting structured data. This makes the values of "currentness" (Section 5.2.3) and "manipulation" (Section 5.2.9) hard to meet. It is also a linear format that cannot be displayed in an alternative way. Individual reading habits (Section 5.2.7) are thus not taken into account. However, hyperlinks to other sections provide a fast way for readers to explore the grammar in an alternative order. Multilingualization (Section 5.2.8) can only be met by providing a second version of the document in another language.

The biggest advantage, however, is the simplicity of the procedure. Linguists are already familiar with digital text editing, most publishers provide books or papers digitally online and, as mentioned before, PDF documents already have a few advantages over

printed books. This is probably the reason why it is already the biggest representative of digital grammars to date.

4.6.2 Hypertext approaches

Another approach is to use hypertext as the medium for a grammar. The grammatical description can be read on a website which functions as the GET. However, the project of creating grammar websites has often been undergone by linguists who also happen to be skillful programmers or by programmers that have been hired by a linguist to create the website from a linear grammar description. Unfortunately, this makes it less suitable for linguists who are not familiar with programming and those who do not have the resources to hire programmers to do it. The first part of a DRG, the GWT, is missing in a hypertext approach.

Websites generally consist of structured hypertext pages that may be styled using stylesheets. Hypertext pages are interconnected documents containing mainly text but also multimedia content such as images, videos, or audio. They are coded in the Hypertext Markup Language (HTML) and can contain hyperlinks to other parts of the text or to other pages. The idea of using hypertext pages as a medium for reference grammars was developed after the rise of the internet in the late nineties of the last century (Zaefferer, 1998). Many theoretical and some practical approaches have since centered around the idea of a digital hypertext grammar (Weber, 2005b; Evans & Dench, 2006; Nordhoff, 2008; Musgrave & Thieberger, 2012; Drude, 2012; Bender et al., 2012).

Hypertext pages have the advantage that they allow for a non-linear structure (see Section 5.3.7). The page sizes are not defined by the format and links between the pages can open alternative paths to explore the grammar. Other advantages can be created by adding functionalities to the pages or to elements within the page (Drude, 2012, p. 162). For instance, audio and video recordings can be embedded and a lexical, terminological, and text database can be linked to the grammar. Possible functionalities will be discussed in detail in Section 5.2. New challenges that arise with these functionalities will be discussed in Section 5.3.

Two examples of grammar websites that use hypertext are “The Arapesh Grammar and Digital Language Archive”[6] (Dobrin, n.d.) and “The Interactive East Cree Reference Grammar”[7] (Junker, 2000). Both projects are websites that are still maintained and contain the grammatical description next to other language- and culture-related resources. While the Interactive East Cree Reference Grammar consists of two versions of a description of East Cree’s morphosyntax (one for the Northern and one for the Southern Dialect) and two chapters on spelling and the sounds of East Cree, the Arapesh Grammar consists of one chapter on phonology and one on morphosyntax, which is still under construction.

[6]Linguistic Editor: Lise M. Dobrin, Technical Editor: Daniel Pitti, Maintenance: Institute for Advanced Technology in the Humanities, University of Virginia

[7]A project of EastCree.org, Editor-in-Chief: Marie-Odile Junker, Interface Design and Development: Delasie Torkornoo

Both grammars use functionalities, such as hyperlinks and audio embedding for lexical and textual examples, and render the grammar entries as distinct hypertext pages. A text-search is only provided for the East Cree grammar.

4.6.3 CMS approaches

As mentioned before, websites can be a suitable means of presenting grammatical descriptions but require programming skills, which are not necessarily common among linguists. A GWT is thus required. While digital text documents can easily be written using a common text editor, adding content to a website requires different tools. Nordhoff (2008, pp. 309-310) and Drude (2012, p. 167) both proposed the use of a content management system (CMS) as a GWT. A CMS is a tool that has been specifically designed to separate the work of web-designers/web-developers from the work of content-creators. The content of a website can be edited without editing the hypertext pages themselves or changing their presentation. After a website is created and the style is defined, grammar authors can add the content of the grammar and modify it using a user-friendly editor within the CMS. The website, however, still has to be created, styled, hosted, and maintained by professionals. A solution would be a platform that is maintained by one central institution and can be accessed by the grammar authors to edit their grammar. Of course, creating and maintaining this kind of infrastructure requires an institution that has a commercial or academic interest in publishing multiple grammars. Nordhoff (2008) proposed and developed such a grammar authoring platform. Unfortunately, the platform has never been backed by a stable institution or organization which is why it has not been maintained and cannot be accessed anymore.

Nordhoff's approach used a special type of CMS, a wiki. Wikis are websites where information is gathered collaboratively by multiple users. Different texts can be read and edited by users and structured in a non-linear way, for example, in a tree structure. Wikis often have features such as a versioning system, where old versions and their differences to the current version can be seen, or access control, where different users can be given individual rights to read or edit texts.

The advantages of CMSs for grammar writing have been listed by Drude (2012, p. 168). He mentions that CMS software (1) already exists and does not have to be developed; (2) usually has version control; (3) allows for collaborative authoring; (4) allows for the distribution of access and editing-rights; and (5) allows the implementation of navigation-tools and full-text-searches. On the other hand, he lists the following challenges (Drude, 2012, p. 168): CMS software (1) does not include functionalities such as a connection to a lexical, terminology, or text database; (2) does not allow for semantic markup (Section 5.3.1); (3) does not easily allow for a linear structure that could guide the reader through the layout of the grammar and also does not allow for linear print-outs.

Nordhoff (2008, pp. 311-318) mentions another software that is similar to a CMSs software: the Fieldwork Language Explorer (FLEx). Its main goal is to facilitate field-

work. It lets users compile an annotated lexicon and provides parsers for semi-automated glossing of texts, among other functionalities. Using the lexical data, grammatical categories, morphological templates, and rules, it is able to generate a morphological sketch of the target language. Although the sketch is not recommended for publishing since it requires further editing, it can be the base of a grammatical description. Nordhoff does not propose it as a GWT per se but uses it as an example, as it meets several of his proposed values of "collaboration" (Section 5.2.6), "manipulation" (Section 5.2.9), "ease of finding" (Section 5.1.1) and "completeness" (Section 5.1.6).

4.6.4 Microscopic and macroscopic approaches

Nordhoff (2008, p. 319) distinguishes what he calls "microscopic" from "macroscopic" approaches to the standardization of grammar. Most suggested formats (Section 6.1), except for the Semantic Web approach by Good (2012), attempt to standardize grammatical descriptions and their formal parts. This is a form-oriented approach that Nordhoff (2008) calls a "macroscopic standard". This approach serves human grammar readers who wish to retrieve information they can find in specific grammar entries. A macroscopic standard enables search queries for specific parts of the grammar.

The other approach aims at the standardization of parts of the language itself. It formalizes language phenomena and is closely related to language engineering. The Semantic Web approach described in Section 6.1.7, for instance, attempts to represent grammatical phenomena and their relations as a directed, labeled graph. This is what Nordhoff (2008) refers to as a "microscopic standard". This approach is primarily directed at the machine-driven processing of language. The grammatical structure of languages could be compared on a much larger scale enhanced by computational tools. Typologists could also enhance their research, for instance, by inferring implicational universals from a cross-linguistic database with a microscopic approach.

Besides the Semantic Web approach, there are other microscopic approaches, such as the attempt to create a Cross-linguistic Reference Grammar database (Zaefferer, 1998; Peterson, 2005; Zaefferer, 2006), which aimed at the formalization of language structures across languages for comparison. This kind of approach relies heavily on the existence of cross-linguistic categories, which has long been a topic of debate (e.g. Haspelmath, 2007; Cristofaro, 2009).

This project will follow a macroscopic approach, as it aims at the digitization of grammatical descriptions and not of languages themselves.

4.6.5 Cross-linguistic approaches

One common reason for the standardization and digitization of grammars is improved comparability. Typologists and other comparative linguists could be able to query through multiple grammatical descriptions and compare language structures much more efficiently. Besides the Cross-linguistic Reference Grammar database mentioned in the pre-

vious section, which aimed at a very fine-grained formalization of language structures, there is also a macroscopic approach to a cross-linguistic grammar platform: the Enhanced Electronic Grammars by the publisher De Gruyter Mouton[8]. This database is fed by several existing grammars published by De Gruyter Mouton. As the publisher enforces a uniform styling of certain structural elements of a grammar (text in target language, interlinear examples, etc.) these elements can be extracted from existing PDFs. However, this method of mapping the style to an element type is rather vague and error-prone. One-to-one-mapping is not always possible, as often one style is attributed to different element types. The "enhanced" PDFs of the grammar chapters now allow for specialized search queries across all grammatical descriptions. This means the user can search a keyword in the examples, in the prose, in the headings, etc. Moreover, these queries can be connected. Additionally, the results can be filtered by region or language family. A missing KWIC view for search results, however, limits the usefulness of these queries. The search results only bring up whole chapters in which the keyword appears. The reader needs to specifically search for the keyword within that chapter. An additional functionality is that abbreviations reveal their expanded term when hovering the mouse over it. As the content of the database only contains retrodigitized grammars, there is no embedding of multimedia or linking to metadata. This database presents an approach to utilize the existing paper grammar format and enhance it by distinguishing differently formatted elements and developing a GET. While this approach is relatively simple, it does without some of the functionalities of a digital medium (e.g., multimedia) and relies on an error-prone method (style-to-function mapping).

[8]retrieved 2020-12-03, from `https://db.degruyter.com/databasecontent?dbid=eeg&dbsource=%2Fdb%2Feeg` – the database requires access rights

Chapter 5

Paper grammars vs. digital grammars

With the rise of digitization, many analog formats have been transferred into the digital sphere and now face their digital equivalent. These two versions of a format do not necessarily need to compete as they might not serve the same set of users or fulfill the same set of functions. Very often, however, digital formats are designed to replace their analog equivalent over time. It is therefore important to investigate what is valued in the analog grammar format and what can be transferred to the digital grammar format. This chapter aims at a comparison between paper grammars and what I envision to be a DRG.

As a matter of fact, grammars can differ quite substantially and there are often two or more opposing opinions on how to write a grammar. Usually, none of them can easily be invalidated. When proposing a digital grammar format, one must consider multiple opinions and provide different ways to structure a grammar. This chapter lists different kinds of values that may oppose each other as they represent different groups of people with different interests. Both a model of grammars (Section 6.3) and values existent in grammar writing will be the prerequisites of the digitization of grammars.

While Section 5.1 lists values that have been expressed in the literature about traditional grammaticography, Section 5.2 adds new values that became more important with the advent of digital grammars. These latter values might also be met by paper grammars but often lack established conventions. Either way, the implementation of these values differs vastly in paper grammars and DRGs. Both sections list values and maxims that can be followed if the value is shared and discusses possible implementations of the maxims in paper grammars and DRGs. The implementation of these maxims in DRGs requires specific properties or features. Those properties and features of DRGs that entail other challenges are discussed further in Section 5.3 and solutions are proposed, if possible. Finally, Section 5.4 combines many of the proposed features and properties into an exemplary concept of DRGs, where possible interaction of those features is presented.

In order to develop and evaluate new functionalities of DRGs, it is important to exam-

ine the needs of grammar authors and writers. The field of metagrammaticography has produced many opinions on what makes a grammar a good grammar and digital metagrammaticography has produced some opinions on what would be desirable in grammars beyond the paper format. These opinions have been extracted as values that can be used to create a the digital data format. Inspired by the presentation in Bird and Simons (2003) and Nordhoff (2008), these values will be listed along with related maxims. These maxims are concrete guidelines that need to be followed if the value is shared. As the proposed digital grammar format in Chapter 6 is meant to be used by as many users as possible, it aims at respecting as many values as possible. Yet, some values oppose each other. In these cases, the values will be evaluated and the most effective and feasible solution will be preferred.

As there are too many values concerning grammaticography for all to be included here, only those that have a direct impact on the format structure or the functionalities of a GWT or GET have been chosen. As the content of the grammar should not be affected by the format's structure, those values concerning the content of a grammar will only be included if they require the format to be more flexible. Moreover, other values concerning the content will also be mentioned if those values are potentially affected by new features of a DRG, discussed in Section 5.3.

Nordhoff (2008) has already given a comprehensive list of values relating to digital grammaticography, and the list presented here is based on the discussion in that paper. However, the list presented in this paper has also been modified and extended to include other values found in the literature in order to be as relevant and updated as possible for further format deliberation in Chapter 6.

5.1 Existing values

At first, paper grammars are examined to isolate what is valued in the analog format. This section lists all values of paper grammars that are relevant to the structure of a reference grammar. Each value will be evaluated with regard to how the value can be transferred to a digital format.

5.1.1 Ease of information retrieval

"We value ease and speed of retrieving the information needed." (Nordhoff, 2008, p. 303)

Readers have different intentions when reading a grammar and therefore read grammars differently. They decide whether they want to take a full path through the grammar or whether they want to pick particular information from the grammar. Readers with the latter intention must be able to navigate easily and quickly through a grammar to access the desired information. This is affected by the grammar's structure and the existence of navigation tools as described in Section 6.3.1.1.

Maxim 1 *Provide a table of contents and index (Weber, 2005b, p. 432; Noonan, 2005, p. 355; Cristofaro, 2006, p. 147; Pawley, 2014, p. 19).*

Easy and fast navigation will also be essential in digital grammaticography. In fact, it is one of the most important arguments for why grammars could benefit from a digital format. This is because a digital medium allows for full text search, more complex and specified search queries, and hyperlinks, which all increase the ease of information retrieval drastically. A table of contents (TOC) and an index, however, remain important navigation tools, as they enable different modes of navigation (Section 6.3.1.1). The implementation of the different navigation tools will be discussed in Section 7.3.4.

In order to follow this maxim, the grammar format needs to

- enable a nested structure of grammar entries so that a TOC can be generated
- provide a way to mark indexes for each section
- link cross-references to the referenced elements
- implement semantic markup in order to allow for specified search queries (Section 6.5.3.3)

A second maxim relates the ease of information retrieval with a familiar structure. The most accessible structure for grammar users is one that they are already familiar with. A grammar should therefore have a structure that is similar to others.

Maxim 2 *Provide a structure that is familiar to the reader! (Nordhoff, 2008, p. 305)*

A digital grammar format will separate the content from its presentation. The presentation of the grammatical description will be the task of the GET. Uniform formatting by the GET can help build familiarity. The structuring of the content, which can also provide familiarity, however, is still in the hands of the author.

5.1.2 Guidance

We value a guided path through a grammatical description (Nordhoff, 2008, p. 305)

As a paper grammar is a linear work with a beginning and an end, it is up to the author how they guide the reader through the grammar. In order to create a coherent story that helps the reader understand the structure of a language, the data may be presented along a guided path. The following maxim applies to everyone that values guidance in a grammatical description.

Maxim 3 *Structure the data in a way you consider the best to guide the reader through the description (Rice, 2005, p. 401).*

Although this value mainly pertains to the content and order of a grammar, the possibility to create a guided path also depends on the structure of a grammar. A DRG needs to provide ways for an author to guide the reader through the description, if this value is shared. This can be especially problematic if linearity is suspended in a DRG (Section 5.3.7). As an opposing view, one can also value the reader's individual reading habit and the possibility of defying the guided path. This value may conflict with the value "individual reading habits" (Section 5.2.7).

5.1.3 Coherence

WE VALUE A COHERENT PRESENTATION OF THE DATA.

A paper grammar is a completed, linear study. The different parts should therefore have a coherent appearance, as cross-referencing is heavily used. The first aspect of a coherent presentation is the consistent use of terminology. Consistent terminology is indispensable to any kind of scientific work. If not applied, readers might get confused or miss links between related phenomena.

Maxim 4 *Use consistent terminology throughout the grammar (Good, 2012, p. 19).*

While authors themselves are responsible for managing their terminology in the paper grammar format, the GWT of a DRG could provide authors with a terminology management tool that can facilitate the consistent use of different terms. Good (2012, p. 20) points out that there is already a substantial amount of research on terminology management and how it can be enhanced, but it is rather uncommon in descriptive linguistics. A terminology management tool could be used to structure an ontology of terms found within the grammar or it can be used to import an existing ontology such as the GOLD ontology for linguistic terms. If used in the grammatical description, a term can be tagged as a "term" and linked to its entry in the terminology database where its explanation can be found. Further details on how a terminology management tool could be implemented are given in Section 5.4.3.2.

The second aspect of a coherent data presentation is the consistency of linguistic analyses. If two analyses applied in a grammar are contradictory, even if situated in distinct linguistic levels, the credibility of the whole grammar decreases.

Maxim 5 *Make the analyses of different grammatical phenomena coherent (Good, 2012, p. 19).*

This aspect is purely concerned with a grammar's content. However, it is mentioned here as it needs to be considered when conceptualizing distributed authorship (Section 5.3.2) as multiple authors may propose different analyses of a certain phenomenon, which would in turn affect the credibility of the presentation..

5.1.4 Comprehensibility

WE VALUE A PRESENTATION OF THE DATA THAT IS EASY TO UNDERSTAND.

This value demands that a grammar is not overly complicated. Besides the avoidance of formalisms and the use of accessible language, which are both aspects that pertain to the content of a grammar, this value also has a structural aspect. In order to guarantee the comprehensibility of examples, paper grammars make use of interlinear morpheme glossing. Utterances in the target language are presented as IGT. IGT is a format that matches not only lexical but also grammatical meaning with a particular morpheme and allows the reader to understand the grammatical phenomena present in a sentence without having to memorize all structures described throughout the grammar as a whole. This method, in addition to mere translation, has become an indispensable part of grammar writing.

Maxim 6 *Provide interlinear morpheme glossing and translation for all examples (Noonan, 2005, p. 355; Rice, 2005, p. 408).*

In order to follow this maxim, a DRG needs to provide a digital IGT format and a way for authors to create interlinear examples. These functionalities are discussed in Section 5.4.3.4.

5.1.5 Accountability

"WE VALUE THE APPLICATION OF THE SCIENTIFIC METHOD." (Nordhoff, 2008, p. 299)

The scientific method in linguistics requires the author to provide the previous steps of every analysis undertaken. The reader must be able to trace back every analysis up to an actual utterance in the language described. In grammars, this is usually done by providing interlinear examples (Section 6.3.2.2).

As exemplar data is carefully selected to illustrate a phenomenon in the best didactic way, it is usually very one-sided. Several diverse examples should be given in order to guarantee that the validity of an example can be checked (Nordhoff, 2008, p. 299).

Maxim 7 *Provide various sources of a phenomenon (Noonan, 2005, p. 355; Rice, 2005, p. 395).*

Paper grammars usually use interlinear examples to provide sources of a phenomenon. Although a larger number of sources would be helpful, too many examples impede the flow of the prose as well as substantially increase the physical size of the grammar itself. Grammar authors thus often adhere to the habit of providing one example per statement made about a particular phenomenon. Further examples can then be searched in an archive deposit or a published text collection. Another method of providing additional sources for a phenomenon is to include an example index that links to other examples

of the same phenomenon within the description[1]. A DRG could utilize this principle and facilitate the provision of more examples by establishing an example database. Examples used within the description could be indexed based on the phenomenon they illustrate. Readers could then be directed to other related examples or could actively search for an example that illustrates a certain phenomenon. These functionalities are explained in detail in Section 5.4.3.4.

5.1.6 Completeness

WE VALUE A COMPREHENSIVE GRAMMAR.

Although it is often postulated that a grammatical description must be complete or comprehensive, it is in fact hard to model completeness. As discussed in Section 4.3.1, maximal outline templates, such as the LDS-questionnaire, can be used for this undertaking. I will, however, use the more abstract concept of Basic Linguistic Theory (BLT) for this maxim as it provides an extensible ontology of concepts and is recommended by many scholars (Noonan, 2005, p. 354; Rice, 2005, p. 403; Good, 2012, p. 14; Pawley, 2014, p. 9). There is understood to be a consensus among descriptive linguists regarding the most commonly used and most widely understood terms in this field, many of which originated in traditional grammar studies.

Maxim 8 *Provide description of all aspects central to Basic Linguistic Theory at the time of publication.*

As the possibilities of language are essentially infinite, and structures within those languages are in constant change, completeness is not so much an unattainable goal but a direction. Payne (2005a, 2014) recommends aiming for the description of only conventionalized patterns among a speech community. This, however, is also a very vague instruction as it is not easy to define "conventionalized", especially because conventionalized patterns are dependent on the speech community's registers.

While the notion of completeness is already hard to grasp in paper grammars, it might be even harder in grammars that can be easily extended or grown over time in the sense of Weber (2005b). DRGs have the ability to grow over time and can be extended at any time. It is all the more important to keep this value in mind when implementing extensibility (Section 5.3.3).

5.1.7 Longevity

WE VALUE A GRAMMAR THAT WILL BE ACCESSIBLE IN THE FUTURE.

A grammar will only be useful for future generations if the format endures time. Paper grammars use books as a medium and libraries as a means to ensure a long life for grammatical descriptions.

[1]This has, for instance, been done by Haspelmath (1993, pp. 530-538) in his Grammar of Lezgian (discussed in Section 6.3.2.2).

Maxim 9 *Provide the grammar in a format that is archivable.*

Digital formats, on the other hand, are less likely to endure decades of technology changes. A format for DRGs should thus be based on existing formats that have already been in use for a while and are accepted by a large community of users. Moreover, DRGs need to be integrated into an infrastructure that ensures the maintenance of the application and the archiving of its content. While the advantages and disadvantages of a digital format are discussed in Section 5.3.1, a suitable infrastructure is proposed in Chapter 8

5.1.8 Persistence

WE VALUE A GRAMMAR THAT DOES NOT CHANGE OVER TIME.

Scientific work has to be citable and therefore not change over time. Paper grammars do not permit fast changes and ensure that the medium persists into the future.

Maxim 10 *Do not change the published resource to enable longterm referencing (Nordhoff, 2008, p. 306).*

As DRGs could be expendable (Section 5.3.3), the value of persistence needs to be considered, as expendable DRGs need a method of consistent citing. How to ensure citability will be discussed in Section 5.3.6.

5.2 Limits to paper grammars

Having listed relevant existing values for analog grammaticography in the previous section, this section presents values that can hardly be met by paper grammars or that are met much easier by DRGs. This section discusses the limits of paper grammars and possible improvements through DRGs. Yet, it is important to note that many of these values could be realized in paper grammars but merely lack a widely-used convention. DRGs, on the other hand, also have limits, as they face new challenges. Those challenges, possible solutions, and limits are the topic of Section 5.3

5.2.1 Accountability

This value has already been discussed in Section 5.1.5 but will be expanded upon by adding new aspects with new maxims. The first aspect pertains to the verifiability of arguments through examples. Transcribed and annotated utterances used as examples in paper grammars are far from being primary data that can be used to verify the arguments. The transcription and annotation has already involved a fair amount of analysis, which is why audio or, even better, video recordings of the actual utterance would greatly improve the scientific quality of a grammar.

Maxim 11 *Provide an actual utterance illustrating every phenomenon described (Nordhoff, 2008, p. 299).*

While paper grammars can provide references to the actual recording, the process of retrieving them and finding the right segment possibly discourages most readers from verifying a faulty transcription, for instance. The presentation of multimedia in paper grammars is simply restricted by the book format. DRGs, on the other hand, could establish a link to the actual recordings in a documentation deposit and embed the right segment of a recording along with the example. The reader would only require one click to play the actual utterance. The embedding of multimedia along with the examples is discussed in Section 5.4.3.4.

The second aspect is also concerned with examples. Language utterances depend on various factors, including properties of the speaker, the situation, etc. This metadata should be provided for every example in order to examine possible effects on the utterance.

Maxim 12 *Provide context for every example (Weber, 2005a, p. 450).*

Two types of context can be distinguished: The example's metadata and the surrounding text of an example. Although it is valued more and more, an example's metadata is usually not found in paper grammars, as it is time-consuming to produce, slows down the flow of the reading, and cannot be hidden. DRGs could access an utterance's metadata through a link to a documentation deposit. This way, existing metadata can be used and less work is required for the author. The surrounding text of an example might be helpful to readers as they might want to look for other information that can be found in the example's environment. Printing the example's textual environment in a paper grammar would be too excessive, but presenting it on demand in a DRG could be a benefit for some readers. As metadata and the surrounding text of the example are not relevant to all readers and might distract other readers' attention, it could be displayed only on demand by clicking on a button below the example.

A third aspect is concerned with metadata of the grammar entry itself. Since grammar entries can be seen as a a resource for information, readers can benefit from accessing metadata before starting to read the entire entry (Bird & Simons, 2003, p. 576). The only metadata of grammar entries in paper grammars that can be accessed before reading the entry itself are the heading, the position within the grammar, and the indexes. The heading is displayed in the TOC and already indicates the topic of the entry. The numerical identifier and the location within the TOC indicate whether the topic is a sub-topic of a higher entry or whether it has sub-topics itself. Indexes are keywords that attribute (sub-)topics to grammar entries.

A DRG could make use of these three metadata types but also add other metadata types. Digital resources can be enriched by "tags". These are usually key-value pairs that label and add a specific keyword to the resource. A heading, for example, could be represented as a tag with the key "Heading" and the value "Phrase types".

Maxim 13 *Provide metadata for every grammar entry.*

An additional metadata type proposed for grammaticography is "quality assessment" (e.g., Nordhoff, 2008, p. 317). Field linguists have a variety of tasks to manage (Grinevald, 2003, p. 60), only one of them is the task of writing a grammar. Among the various topics covered by a grammatical description, there are usually only a few that the author is an expert in. This is inevitable, even if we raise the number of experts by means of distributed authorship. The reader, however, could benefit from the knowledge of whether a phenomenon has been analyzed by an expert of the corresponding field or not. Quality assessment could transmit this information by indicating the confidence an author has in the present analysis. This is sometimes stated in paper grammars through indications such as "more research is needed" or "it looks like".

A more specific assessment could be achieved in DRGs through a tag. Such a tag could be the quality stage of an entry indicated by values such as "note", "sketch", "draft", or "peer reviewed" (Nordhoff, 2008, p. 317). This tag could theoretically also be printed at a central spot of a grammar entry in paper grammars. However, such a convention has not been established. In expandable or updatable DRGs, however, this would be more important, as early stages of a grammar entry could be published. The indicated quality stage would then inform the reader about how far the research has advanced (Section 5.3.3).

Likewise, readers could also benefit from the knowledge of whether a grammar entry is relevant for them. Relevance of an entry is dependent on the reader's intentions and can be indicated through different kinds of metadata. The content of an entry is already indicated through a header and indexes. They can be used if readers are searching for specific information they know how to label. However, every language has its own categories and is often described in its own terms. It could therefore be important to mark the cross-linguistic relevance and the relevance within the language (Nordhoff, 2008, p. 317).

While cross-linguistic relevance indicates how common a specific phenomenon is across languages, the relevance within the language indicates whether a language structure is central or rather peripheral. However, assessing these criteria bears some challenges. Firstly, how can cross-linguistic relevance be assessed? What makes a phenomenon a common phenomenon across languages? Authors could just use different or obscure terminology to make a phenomenon seem like it is rare, although it is not. Secondly, how could an author assess whether a phenomenon is central or peripheral? Does the frequency have to be measured or does it depend on the author's analysis? Both categories seem to be very impressionistic unless quantitative research is conducted. Due to time constraints, however, most grammar authors might not want to make this effort. Relevance of an entry can also be marked in paper grammars and is sometimes mentioned in the prose. Establishing relevance as a metadata category in DRGs, however, converts this information into machine-readable language that could then by used by external applications. Yet, in order to have reliable metadata categories, conventions would have to

be established or made more rigid.

Relative importance of an entry and the stage of the quality of its content are only two metadata categories that have been proposed for grammaticography. If they are defined well and made reliable through consistent conventions, they could help the reader find the right grammatical entry. If they are displayed in the TOC or other navigation tools, readers could assess the relative importance of an entry before starting to read. This is especially important for a non-linear approach to grammar writing (Section 5.3.7).

5.2.2 Accessibility

WE VALUE A GRAMMAR THAT IS ACCESSIBLE ANYWHERE.

As a product of scientific research and a medium that contains linguistic knowledge, grammars should be as accessible as possible, independent of the location of the reader. Unlimited access, however, is unattainable as no existing medium is free from locational constraints.

Maxim 14 *Make the grammar accessible to as many people and at as many locations as possible (Bird & Simons, 2003, p. 566).*

Access to paper grammars and access to DRGs, however, vastly differ. Paper grammars can usually be accessed from certain libraries, often situated in the Global North. DRGs, on the other hand, could be distributed through the internet. This, however, requires a device to access the internet, such as a computer, a tablet, or a smartphone. While computers or tablet computers would ensure an easy operability of a DRG, the access to computers is far from being global. The use of smartphones, on the other hand, has increased drastically, even in remote areas. However, accessing a DRG from a smartphone requires a responsive design that is oriented at small display screens and touch interactions. Internet connectivity is another constraining factor that is especially problematic in remote areas and countries of the Global South. While access to the internet is not given everywhere around the globe, many people may still have access at certain locations they frequent. A DRG would be greatly improved if most of the functionalities could also be used offline after it has been retrieved once through the internet.

5.2.3 Currentness

WE VALUE STATE-OF-THE-ART ANALYSES AND SCIENTIFIC PROGRESS (Nordhoff, 2008, p. 299).

There are several reasons why a grammar's content could become outdated. Recent fieldwork or other data retrieval might enable the author to add new parts to the grammatical description or to revise existing analyses. On the other hand, linguistic theory is constantly evolving and the latest findings might motivate the author to analyze certain phenomena differently. Either way, readers could benefit from updated grammars in order to access state-of-the-art analyses.

Maxim 15 *Provide a way to update the content in order to give state-of-the-art analyses (Nordhoff, 2008, p. 300).*

Paper grammars are rather rigid and updating the content would mean to publish a new edition of the book, which is a rather time-consuming process. Yet, publishing a new edition also ensures the quality of the new edition as it usually is peer reviewed again. DRGs, on the other hand, could feature a text editor that enables the content to be modified or expanded over time. This new feature, however, entails other problems as discussed in the following value "history" and in Section 5.3.3.

5.2.4 History

WE VALUE THE DISCLOSURE OF PRIOR ANALYSES (Nordhoff, 2008, p. 300).

The previous maxim, if applied alone, would result in a changing resource, from which previous analyses cannot be retracted. There is value in examining the logic of prior analyses as this leads to a richer understanding of the status of the research at a given point in time. The following maxim demands that scientific progress should comprise not only the current analysis but also prior analyses that influenced the current one.

Maxim 16 *Provide a way to present prior analyses (Noonan, 2005, p. 360; Nordhoff, 2008, p. 300).*

As new editions of paper grammars are produced if the content is revised, all prior analyses exist within different editions. If DRGs can be updated, they need a feature to manage updates as separate versions that can be accessed by the reader. Such a version control/version history feature is discussed in Section 5.3.4

5.2.5 Author-assistance

WE VALUE GUIDANCE AND SUPPORT FOR THE AUTHOR (Nordhoff, 2008, p. 300).

A grammar author's main tasks should be linguistic analysis and the description of a language. Other time-consuming tasks, such as layouting, are forced upon the author. An application where the layout is automized so that the author can concentrate on more essential tasks would be favorable. This should also include the automatic numbering of grammar entries, tables, examples, and figures (Weber, 2005b, p. 432).

Maxim 17 *A grammar author should not have to be concerned with a grammar's layout (Nordhoff, 2008, p. 300).*

While it is necessary for the author of a paper grammar to construct the grammar's layout, a digitally encoded grammatical description can be automatically formatted in a uniform way or even in multiple ways from which a reader can chose. This way, the reader can follow their own reading habits (Section 5.2.7) and chose a layout they are

familiar with (Section 5.1.1). However, it might also be preferred that an author is able to make adjustments to the automated formatting as some layout choices might help the author better bring their point across. For instance, the formatting of interlinear examples can influence the reader's attention (Section 6.3.2.2).

Templates, such as the maximal outline templates described in Section 4.3.1, could act as a guide for the author in terms of structuring the grammar. Moreover, templates, if used by multiple authors, result in familiar structures that improve the readability and comparability of a grammar. Not only can the overall structure of a grammar be modeled with templates, but the internal structure of a grammar entry can be standardized using a template as well. Nordhoff (2008, p. 301) mentions the additional advantage of semantic markup, which could be placed in a grammar entry template (Section 6.5.3.3). This enhances the machine-readability of grammars and facilitates data manipulation by third parties.

Maxim 18 *The grammar author should be provided with templates for the overall structure of the grammar and for the structure of smaller units within a grammar as well (Weber, 2005b, p. 430; Nordhoff, 2008, pp. 300-302).*

In a DRG, the GWT is responsible for author assistance and the provision of templates. While templates for the overall structure (minimal or maximal outline templates) should merely be suggestions and therefore optional, templates for the internal structure of grammar entries need to be a mix of obligatory and optional elements in order to ensure format compliance. The proposed data format (Chapter 6) provides the formal elements that can be filled with content, but most grammar-authors are not familiar with markup languages such as XML used within a format. A GWT should therefore contain an editor that guides the author in encoding the grammatical description in a format compliant manner without confronting them with the markup language. The XLingPaper format is a good example of the use of templates. Using the XMLmind editor, documents can be edited in the XLingPaper format without the knowledge of XML (see Section 6.1.4).

In order to write a new grammar entry, the author could, for instance, use a template for a grammar entry. A template can contain obligatory, recommended, and optional elements. For the grammar entry, the editor could force the author to add a title element, recommend additional keyword elements for the index, and let the author chose from optional elements such as paragraphs, tables, examples, etc. Each template should be designed carefully to separate elements that are indispensable and elements that are important but often forgotten.

Semantic markup can also be ensured through templates. Linguistic units, such as syntactic units, lexemes, or word forms could have their own templates. If templates are used by authors, the computer can differentiate those units and associate them with different functionalities. Drude (2012, p. 173), for example, proposes the following functionalities for those units: a syntactic unit could be clicked to play its linked media file, a lexeme could lead to its lexical entry, and a word form could reveal interlinear glosses for

the morphemes it contains. The use of templates and the process of writing a grammar within a DRG is discussed in Section 5.4.3.1.

5.2.6 Collaboration

WE VALUE MULTIPLE CONTRIBUTIONS TO A SCIENTIFIC STUDY (Nordhoff, 2008, p. 302).

Grammar writing has predominantly been a one-person-task. This leads to a situation in which one person engages in the description of various linguistic phenomena of which the author might not be an expert. While collaboration is equally possible for authors of paper grammars, it could be facilitated through a GWT that lets multiple authors write on the same grammar and manage a common set of terminology, examples, etc., to ensure consistency.

Maxim 19 *Provide a way for many researchers to work on a grammar from different locations (Weber, 2005b, p. 422).*

Even though a DRG might facilitate the collaboration of different authors, there are still a lot of challenges associated with different types of collaboration. Section 5.3.2 elaborates on the types of collaboration, the challenges they entail, and how these challenges can be solved in general, especially with features of a DRG.

As collaboration comes with so many challenges, researchers might also prefer to work alone and share the value "individual work", which opposes the value "collaboration". The value "individual work" has not been included in this list, as the default process of producing a grammar is single authorship. Both formats, paper grammar and DRG, are primarily tailored to the needs of single authors.

5.2.7 Individual reading habits

WE VALUE THE READER'S INDIVIDUAL READING HABITS (Nordhoff, 2008, p. 303).

Every reader is different, as they have different backgrounds, might search for different information, or prefer a particular way of reading a grammar. This is why a grammar should give the reader the freedom to take their own path. This value is contrary to the value of guiding (Section 5.1.2). If both values should be met, the choice has to be left to the reader whether to be guided through the grammar or to explore an individual path.

Maxim 20 *Provide a way for the reader to follow their own path.*

Besides a guided path that is usually chosen by the author through the ordering of the grammar entries, there might be other reasonable ways to read through the grammar. In order to allow the reader to take alternative paths in DRGs, Nordhoff (2008, p. 315) suggests a dynamic approach to the structure of a grammar. The grammar entries are not ordered but belong to certain domains and are tagged with a priority value. This approach and all problems arising from the (non-)linearity of a DRG are discussed in

Section 5.3.7. Paper grammars are confined to linearity and follow a static path chosen by the author. They do, however, also provide alternative paths via cross-referencing. Paper grammars are dependent on a lot of cross-referencing, and this enables readers to follow their own paths and quickly find the relevant section.

Maxim 21 *Provide a way for the reader to access cross-references quickly.*

DRGs can also follow the static approach by providing a default path and providing a lot of cross-references. While finding referenced entries in a paper grammar can be time-consuming, digital grammars could provide hyperlinks that connect two related entries through only one mouse-click. While the static approach does not differ from that of paper grammars, cross-references can be accessed much quicker, as they will be represented with hyperlinks. Hyperlinks can also be used in navigation tools such as the TOC and the index to provide quick access to a particular grammar entry. Another way to quickly search for a topic is through full text search or other search functionalities. These could be equipped with the ability to read query languages or regular expressions to enable more efficient searches (Section 7.3.4).

5.2.8 Multilingualization

WE VALUE THE PROVISION OF SCIENTIFIC WORK IN MULTIPLE LANGUAGES (Nordhoff, 2008, p. 307).

Language documentation is valuable not only to an academic crowd of typologists, but also to other linguists, language activists, and community members. Although I have confined this digitization project to the audience of academic linguists, providing a grammatical description in multiple languages increases its accessibility. Linguists, specialized in a certain area or language family, might benefit from a grammar written in the language of wider use in the area where the language is spoken, as they might be more proficient in it.

Maxim 22 *The grammar writer should be enabled to publish the grammar in multiple languages (Weber, 2005b, p. 433).*

Authors of paper grammars can provide multiple versions of a grammar in different languages by publishing translations of the grammar. DRGs, on the other hand, could assist the author in producing parallel elements of a grammar in different languages. The different versions could be read in one place, allowing the reader to switch the target language easily. In order to make a grammatical description available in multiple languages, the format needs to the format needs to allow for multilingual content. XML, for instance, has the attribute `@lang`, which indicates the content language of an element. Thus, a grammar entry that is represented by the element `<entry>` could have the attribute `@lang` with the value `en` for English to indicate that its content is in English. A second element `<entry lang="es">` could have the same content in Spanish.

A grammar in different metalanguages could be regarded as a grammar with different layers, each metalanguage representing one layer. Likewise, grammars could have other kinds of layers, for instance, parallel content specific to different audiences. A particular grammar entry could then have one instance that describes the phenomenon for linguists and one for language learners. This way, a reader could switch the layer in the event that they are searching for additional information from another perspective. A language learner, for example, could switch to the linguistic layer to look up a linguistic explanation of a particular phenomenon. Another reason for layering a grammar could be the parallel description of target language varieties.

5.2.9 Manipulation

WE VALUE THE DESIRE OF READERS TO REUSE THE DATA (Nordhoff, 2008, p. 307).

As grammars contain a lot of data that could be used for other scientific purposes, it is important for readers to extract and manipulate data. Paper grammars only allow for copying the data, whereas digital grammars could have the possibility to extract selected data in a standardized format in order to manipulate it later. This, in turn, will require standards to attribute data to the original source.

Maxim 23 *Provide a way for the reader to extract and manipulate the data presented in a grammar.*

A GET could provide different kinds of outputs of the whole grammar, parts of the grammar, or data from one of the databases. A linear book version (probably in the PDF format) could be useful in making a print version while the output of the DRG format or other open formats could be given out to extract its data or to incorporate the data into other applications.

5.3 Challenges, solutions, and limits of DRGs

The previous two sections have listed values of grammaticography and have proposed different features and properties of a DRG that could be implemented to meet those values. This section further discusses those features and properties that entail certain challenges or violate other values. If possible, solutions are proposed for each challenge. If not possible, an unsolved challenge will be regarded as a limit of DRGs. The first subsection will not discuss challenges and possibilities of a specific feature of DRGs but of the digital format in general.

5.3.1 A digital format

A grammar is a coherent story about the structure of a language that grabs pieces of information and presents them at the right time in the right way. This has been referenced

earlier as a compromise between a story and a database. Digital formats have the advantage that information can be stored in different layers and separated from the layout. A website, for example, has its content structured using the markup language HTML and separated from the presentation of the content, which is defined using the style sheet language CSS. For grammars, that means a digital format could separate the database from the story. Structured information, such as interlinear examples, dictionary entries, or an ontology of terms could be stored in a structured way that is machine-readable. A prose text could contain the story that references the information in the database. Once the digital grammar is displayed by an application, it can combine the story and the structured information into one coherent layout, which is a familiar aspect that is similar to paper grammars. Storing the information in a machine-readable way not only helps the computer find information to facilitate fast and effective searches, but also allows for data extraction and an alternative, more data-driven layout. For instance, interlinear examples could be stored in an example database or terms in a terminology database. Readers who already have an idea of the language's structure might not need a lot of explanatory text but might want to access paradigms directly, probably to compare or manipulate them. They might want to use the grammar as a database and prefer a data-driven layout.

Developing a digital format for reference grammars also means that reference grammars are going to be standardized to some degree. Standardization does not necessarily mean a rigid mold for a creative product, but could also consist of rather flexible recommendations. In fact, standardization can be situated on a scale between what Nordhoff (2008, p. 319) calls a "rough guideline" and a "law-and-order" approach. Advantages and disadvantages of a digital format depend on the approach of standardization. A rough guideline approach allows for flexibility and creativity and usually entails a rather flat learning curve. A disadvantage of this approach is the risk of inconsistency, especially across different grammatical descriptions produced by different authors. The law-and-order approach, on the other hand, does not leave room for much flexibility and users might be disinclined to use such a format. Yet, it enhances the machine-readability and increases the comparability of grammatical descriptions. An extreme form of the law-and-order principle are the microscopic approaches mentioned in Section 4.6.4. This thesis, however, uses an approach that is closer to a rough guideline, as its main goal is to be acceptable to as many users as possible in order to become a widely used standard. The proposed format will provide the formal elements that are also present in paper grammars and will be enhanced by optional semantic markup.

In this thesis, "semantic markup" refers to the provision of elements within a grammar format that do not represent superficial parts of the grammar such as paragraphs, tables, or examples. Those elements could be linguistic units such as phonemes, morphemes, lexical entries, or grammatical categories. Instead of representing these units in the format as a formal unit (for example, to surround terms with quotation marks), they will be represented according to their function. This has the advantage that linguistic

units and grammatical categories can be distinguished by the computer and taken into account when searching for information within the grammatical description. This could, for instance, be important if a reader searches for peculiarities related to the phoneme /p/. If occurrences of the phoneme /p/ do not receive semantic markup within the description, the search result for the letter "P" would contain hundreds of instances where words contain the letter "P". A specific search for the phoneme /p/ is only possible if all instances of the phoneme /p/ within the text are marked as phonemes. Semantic markup holds even more value when vast amounts of data are compared automatically. As semantic markup is more time-consuming, even if a GWT assists the author, it should not be enforced but only be recommended. An author can decide whether they want to invest more time to enhance the grammatical description with semantic markup or not.

Another important advantage of a digital format is that it can easily be accessed from a great number of locations throughout the world if disseminated through the internet (Section 5.2.2). On the other hand, preserving digital documents or applications on the internet for a long time requires a different process than that of paper grammars. Publishing a grammatical description online means a document is stored on a server somewhere in the world. As this particular server could experience damage, there should be backups stored on other servers. In the hypothetical case that a GET/GWT environment is maintained by an institution such as a digital publisher, an archive, or a university, the procedure of producing backups is probably a common routine (Nordhoff, 2008, p. 314). Yet, whenever an author edits an unpublished version of a grammar using a GWT, it needs to schedule regular automatic backups to prevent data loss. If the GWT is used offline, a backup can only be stored on the user's computer, but in case that the computer gets lost or irreparably damaged, data loss cannot be prevented.

Backup WE VALUE REGULAR AUTOMATED BACKUPS (Nordhoff, 2008, p. 302).

Maxim 24 *A DRG should provide regular automated backups of the published grammatical description and of its unpublished versions (Weber, 2005b, p. 434; Nordhoff, 2008, p. 314).*

DRG applications need to be maintained, which is labor- and cost-intensive. If applications to read or write a specific grammar format cease to exist, the format itself will not be useful anymore, as it cannot be displayed or modified. The utilization of a widely-used mark-up language, such as XML, guarantees that the infrastructure for extracting data from XML-based documents will still exist in the future.

Tangibility WE VALUE THE HAPTIC-VISUAL EXPERIENCE OF READING A GRAMMAR (Nordhoff, 2008, p. 307).

For some users, another deficit of a digital format is the fact that DRGs cannot be held in the hand like a book. Nordhoff (2008, p. 307) describes holding a physical copy of a grammar as a "holistic experience of accessing the GD [grammatical description] with all senses".

Maxim 25 *Provide a grammar that can be held in the hand (Nordhoff, 2008, p. 307).*

It is not possible to follow this maxim with a DRG, which is why "tangibility" could be regarded as a limit of DRGs. However, the provision of different outputs (Section 5.2.9) could also include a printable version of the grammar that can be held in the hand. Yet, an output printed at home or in the office might not satisfy all readers that value a grammar that can be held in the hand.

Many advantages and disadvantages of a DRG are related to the different features and properties that can be provided in a digital setting. These functionalities and properties, the challenges they entail, and how they can possibly be overcome are discussed in the following sections.

5.3.2 Distributed authorship

As mentioned in Section 5.2.6, when discussing the value "collaboration", collaboration between different authors is possible in paper grammars and DRGs. However, a digital format and a GWT open new possibilities to facilitate the process and the coordination between multiple authors. However, collaboration between authors can take on various forms that entail different challenges. Three types of collaboration will be distinguished here:

1. Co-authoring of a grammar monograph
2. Co-authoring with a principle grammar writer and other contributors
3. Grammar as a collection of independently publishable parts

The first type of collaboration aims at the production of one coherent grammar monograph written by two or multiple more or less equivalent authors. The biggest challenge of this attempt is to not interfere with the value "coherence". As the co-authors attempt to produce one coherent story, they need constant communication about used terminology, the analysis of all phenomena, the order of topics and, more generally, how the story will be told. This task can consume large amounts of time and requires a lot of perseverance. In some cases, it might even be impossible due to opposing opinions.

The problem of coherence could be moderated by the second type of collaboration. It also aims at the production of one coherent grammar monograph but with only one main grammar writer and other contributors. This division of labor guarantees a coherent story that is only written by one person. Other contributors might help with the processing of data, single analyses or simply the provision of knowledge. Weber (2005b, p. 420), for instance, mentions the possibility of co-authoring with a retired field-linguist, who still knows a lot about the language that might otherwise be lost. A problem with this approach might be the attribution of contributors to what they contributed. The looser a contribution is (for example, personal conversation) the less clear it becomes what exactly was contributed.

The third type of collaboration proposes a clear attribution of each author to what they have contributed. It does not aim at the production of a monograph but at a collection of independent, publishable units that together describe the structure of a language. As coherence among the independent units is only marginal, less coordination is required and collaboration becomes easier. However, it might be very difficult for readers to understand the structure of a language as a whole when there is not one coherent story about it. Good (2012, p. 13) elucidates on the concept of distributed authorship, where multiple authors contribute descriptions of a language in the area they are specialized in. What he envisions is an atomization of research results with smaller units that can be published independently. However, he combines this vision with the use of a microscopic standard (Section 4.6.4), where facts of the language are encoded rather than the elements of a grammatical description. This vision does not deal with a coherent story or with transitions between topics. The (dis)advantages of this approach and the use of the semantic web are discussed in Section 6.1.7. As mentioned earlier, this project will follow a macroscopic approach and aims at the encoding of coherent grammatical description. Type three is therefore excluded from further discussion.

A DRG could offer some help for the coordination between authors. As a GWT could be accessible online, different users should be able to access it (Weber, 2005b, p. 422). A user management tool could be implemented where the main author could grant editing rights to other users to modify the whole grammar or particular entries. Different authors could thus work on the same text and see all modifications of their co-authors. Good (2012, p. 19) also suggests a terminology management tool to organize the used terms and the use of grammar engineering systems to see if different analyses are compatible. A GWT that allows for modifications by multiple authors, especially in combination with offline usability, requires a version control feature (Section 5.3.4). Each time an author has modified parts of the grammar, they can decide to share the version with their co-authors or even publish it. Versioning software now appoints the new version to be the current version, while storing the old version. Difficulties may arise with the synchronization of multiple versions from different authors, especially, if they have modified the same part of the grammar (Drude, 2012, p. 105). Existing version control systems, including those used in a CMS, tackle this problem by providing a merge-function. The author of the second of two conflicting versions is asked to compare the two versions when trying to save it online. For each line, they need to decide which elements of the two versions should be transferred to the merged version. Sometimes, this process is complicated and time-consuming, but it can be avoided with an organized procedure and communication between the co-authors. The resulting maxim can be added to the value “author-assistance”.

Author-assistance WE VALUE GUIDANCE AND SUPPORT FOR THE AUTHOR.

Maxim 26 *A GWT should provide a guided process to merge conflicting versions.*

Another challenge that arises with distributed authorship is the attribution of contributions to each contributor. The question who attributed what can also be answered by a version control system. As shown by Nordhoff (2008, p. 313), some version control systems of CMSs also provide the functionality to mark which author has edited which part of an entry. The following maxim can be added to the value "collaboration".

Collaboration WE VALUE MULTIPLE CONTRIBUTIONS TO A SCIENTIFIC STUDY.

Maxim 27 *Indicate which parts of the grammar have been contributed by which author (Nordhoff, 2008, p. 302).*

As shown, none of these approaches is close to being perfected, as all of them face serious obstacles in practice, even with functionalities of DRGs that could improve the process. Yet, the advantages of joining forces and improving the quality of a grammar are promising, which is why new conventions need to be established. Having discussed different aspects and possibilities in distributed authorship, this thesis will not consider this topic any further in the upcoming discussion. It will, however, propose a format that should not impede the process of distributed authorship.

5.3.3 Extensibility and updatability

These two properties can be seen as advantages of DRGs, but they also entail a few challenges. Extensibility means that the grammatical description of a DRG can be expanded at any time. As envisioned by Weber (2005b), a grammar could thus grow over time. While early findings could already be published immediately, the grammar could be expanded every time that new data or new analyses are available. Updatability means that the grammatical description can be updated at any time. In order to follow the maxims of the value "currentness" (Section 5.2.3), grammars should be updatable whenever new data or new findings in linguistic theories require a change in analyses.

As mentioned earlier, these two properties could affect the value "persistence". The solution to this problem is discussed in Section 5.3.6. A further challenge concerns the value "completeness" (Section 5.1.6). As completeness is hard to measure, the value can hardly be met to perfection. However, the feature "extensibility" might tempt authors to publish parts of the grammar and extend it later. Such a grammar is far from complete in the beginning. Yet, an incomplete grammar is still more valuable than no grammar. However, in this case, it would be valuable to mark the quality stage of each grammar entry so that readers have the possibility to extract valuable information in an informed manner[2] (Section 5.2.1). The discussion about completeness is particularly relevant in the field of documentary linguistics, where the amount of documentation confines the

[2]It should be noted that automatic extraction of information through external applications might not pay attention to the quality stage of an entry. New conventions for data extraction that take into account different quality stages need to be established.

possible extent of a grammatical description. Good (2012) calls this completeness in relation to the amount of available documentation "coextensivity".

"Coherence" is another value that could be affected by extensibility and updatability. If a grammatical description grows or changes over time, the risk of producing incoherencies increases. This risk, however, is also given with paper grammars, as those are also products of several months and years. The only difference is that a paper grammar will be proofread and checked for coherence as a whole. Constantly extended and updated grammars would need to be checked for coherence every time they are extended or updated. This is a time-consuming process and might often be skipped for small edits. A helpful tool could be the terminology management tool, as it organizes the used terms.

Peer review is a common standard in publishing to ensure the quality of academic writing. While paper grammars are peer reviewed before they are published, peer reviewing for updatable DRGs raises questions: How will an extended grammar or updated entries be handled? Could single grammar entries be peer reviewed in isolation? The current system of peer reviewing does not seem to be prepared for a constant reviewing process for updated grammar entries, as it is already at its capacity. Moreover, reviewing single grammar entries that are part of a coherent story is probably very ineffective if not impossible. Solutions to the problem of ensuring a high scientific standard for updatable DRGs are yet to be found.

5.3.4 Version control

A GWT that permits updates to the grammatical description also needs to have a version control functionality, where new and old versions can be managed. This is indispensable if multiple authors are allowed to modify the same grammar, as discussed in Section 5.3.2. It is also important to connect a grammatical description to a certain point in time which would enable stable citing. Older published versions need to be accessible in the GET as an old version might have been cited in another source. This functionality is referred to as "version history". A version history functionality in a GET could, for example, be like a filter where the desired version can be selected. A precise implementation of this feature is presented in Section 7.3.4. An advantage of a CMS approach (Section 4.6.3) is that many existing CMSs already provide a version control feature.

5.3.5 Templates

Templates are an effective tool to control the uniformity and the machine-readability of grammatical descriptions (Section 5.2.5), but they also have disadvantages. The criticism for grammar outline templates such as that of the Lingua Descriptive Series (Comrie & Smith, 1977), for instance, often revolves around the metaphor of pressing a language into a mold (Weber, 2005b, p. 430). Nordhoff (2008, p. 302) remarks that templates should not interfere with the author's creativity and Drude (2012, p. 169) suggests the use of templates only for some parts of the grammatical description.

Creativity WE VALUE THE AUTHOR'S CREATIVE FREEDOM

Maxim 28 *A GWT should not impede the author's creativity. (Nordhoff, 2008, p. 312)*

A mix of obligatory, recommended and optional elements, as proposed earlier would be an efficient way to guide the author but to also value their creative freedom.

5.3.6 Citability

How to cite a DRG or parts of it is not obvious. When proposing a digital format for reference grammars, one needs to rethink the questions "How can consistent citing be enabled?" and "How can specific parts of the DRG be cited?".

Consistent citing (Section 5.1.8) of paper grammars is not difficult to achieve. A book is a stable medium where the same data that has been cited can be accessed in the future. Data within DRGs that can be updated and extended, however, might not be the same in the future. If authors are able to constantly edit their grammar entries, the document will not remain in the same state and longterm referencing cannot be ensured. A solution to this problem is the version history feature (Section 5.3.4). With access to all older versions of the grammar, citations that indicate the version number will remain consistent. The practice of indicating the version number, however, needs to be adopted by the readers. A GET thus needs to provide guidance for citing. Ideally, parts of the grammatical description that can be cited will have a "cite" button, which produces a formatted reference to the selected part, including the version number. Additionally, the reference could be provided as an output in one or multiple common bibliography formats in order to help the reader incorporate the data into their preferred referencing tool. If working with different versions, the reader also needs to be able to find a specific version. A page displaying the version history of the entire grammar – including the databases – could help readers select a particular version. Additionally, each grammar entry could have a separate tab for its own version history. This way, readers can trace back changes in the particular entry and see which edits can be attributed to which author.

As DRGs do not necessarily provide a layout with consistent page sizes, the question "How can specific parts of the DRG be cited?" arises. If hypertext pages each represent a grammar entry, very long pages might exist. Cited information, especially when paraphrased, could be hard to find. A solution would be to provide a way to not only cite grammar entries, but also smaller elements such as paragraphs, tables, or examples. While tables and examples usually already have numerical identifiers, paragraphs do not. Displaying these numerical identifiers at the top of each paragraph might be distracting for the readers. Instead, the identifier could only be displayed when clicking an unobtrusive "cite" button (Section 7.3.2). In any case, a reader needs to be able to locate a cited paragraph within the grammar. Therefore, search functionalities need to provide a way of locating paragraphs by means of the numerical identifier.

5.3.7 Linearity

Books as a medium have shaped the structure of grammatical descriptions immensely. A grammatical description needed to be written on multiple pages of identical size, ordered from left to right (or right to left in some non-Latin scripts). In this way, books have forced linearity on grammatical descriptions – to the extent that a grammatical description without linearity is hard to imagine. Although authors of paper grammars have defied linearity by using a complex nested structure of grammar entries and providing a great number of cross-references, a grammatical description bound to a book still constitutes a path from the front cover to the back cover. A digital medium, however, does not force linearity upon grammatical descriptions. This has triggered a discussion about the question of whether linearity belongs to a grammatical description or whether grammatical descriptions would profit from a non-linear structure. This discussion reflects the metaphor of a grammar being a compromise between story and database. While the story represents linearity, the database represents non-linearity. The question is now, "Is a grammatical description a database that has been forced to look like a story? Or has the grammatical description always been a story with traits of a database?" While the first option would demand non-linear DRGs, the latter would demand a linear structure within DRGs. The reality is that this question is impossible to answer. Moreover, readers have been accustomed to reading linear grammars and might not even benefit from a revolutionary break with linearity. Nevertheless, this section will discuss advantages and disadvantages of a non-linear structure.

Whether linear or non-linear, a grammatical description deals with the structure of a language. Regardless of how a language's structure is broken down, its subdivisions can hardly be described in isolation, as they are all interrelated: "The most important point is that a language can only profitably be studied as a whole. One must recognise and distinguish different levels of structural organisation – phonological, morphological, syntactic, semantic, discourse and pragmatic – but each of these continuously interrelates with the others" (Dixon, 1994, p. 229). Both linear and non-linear descriptions rely on connections between grammar entries. The wiki approach mentioned in Section 4.6.3 provides a non-linear structure for multiple entries. It does not necessary have an order, but is a collection of entries that may reference each other and can be searched through a search field. A wiki approach is very suitable for encyclopedias, where specific information is usually searched for and may be found in one entry or multiple related entries. A grammar may be used the same way if a specific phenomenon or a specific function is looked for. It may, however, be less suitable for readers trying to learn about the entire language.

A solution is to provide a "dynamic path" through a non-linear structure. Nordhoff (2008, p. 315) proposes a collection of grammar entries produced in a CMS. Each entry belongs to a grammatical domain[3] and is given a priority value by the author. This way,

[3]Nordhoff (2008, p. 315) proposes fine-grained domains such as "DomainVocalicPhonology", "DomainIntonationalContour", and "DomainVerbalPrefixes".

the desired domain can be searched and relevant entries are displayed in the order of the priority. Other readers might only be interested in one phenomenon and want to know how this phenomenon is integrated in the language's structure. As each entry is part of at least one domain, the domain an entry is in could be browsed to understand the context of a phenomenon. For instance, a reader is interested in the habitual marker of a language. They might have found the respective grammar entry and would now like to know more about its context: What other TAM markers are there? Where are TAM markers located in the sentence? etc. As the grammar entry about the habitual marker is part of the TAM domain, this domain could be browsed or read entirely in order to answer these questions.

This approach, however, seems to renounce the idea of a coherent story through the whole grammar, as readers are asked to create their own path through it. Moreover, it seems to do without transitions between grammar entries. This might also have an impact on how authors write the grammars, as it might be unfamiliar to write grammar entries without a preceding or following one. Authors always need to consider which other grammar entries readers need to have read in order to understand the current entry. In the event that they haven't read those entries, authors might want to redirect readers to those entries. This could be a list of entries at the beginning of an entry that are deemed useful for understanding the current entry if read in advance. However, if almost every entry has such a list, loops might be created. The dynamic path approach seems to need more examples in order to clarify these open questions.

Nordhoff claims that, in contrast to a static path (a linear structure), the dynamic path does have the difficulty that new entries have to be inserted into the path. According to him, these insertions can be difficult, as a new entry might not have a suitable place within the path or errors could occur when inserting the new entry. This should, however, not be a problem. Grammar authors have always had the same difficulties when conceptualizing grammars. Imagine, for example, an author who has already written a substantial amount of their grammar but wants to describe another phenomenon that does not fit perfectly into the grammar. There are a few possibilities to handle this situation. Either the structure is slightly changed, a super-heading is extended to facilitate the inclusion of the new entry, or the text of the entry needs to be inserted into another grammar entry. This problem has probably existed and been solved since the creation of the first grammar.

As for the technical problems of inserting a new entry to an ordered list of entries, when creating a new entry in a GWT, the author needs to choose a place in the existing structure. To facilitate this process, a simple "drag and drop" visualization could be implemented. Most authors are familiar with this process from popular WYSIWYG editors. The GWT could do the rest and automatically insert the new entry with all its elements in the desired place within the DRG file. The only problem are numerical identifiers of the grammar entries. If it is the case that entries are numbered and the numbers are used as to reference a certain part of the grammar, new grammar entries cause a renumbering

of a certain amount of entries. This problem is of the same nature as the one discussed earlier in Section 5.3.6. If entries are cited together with the version number, the specific entry can easily be found in the indicated version.

The lack of linearity can affect the value "guiding". The advantages of a guided path through the grammar have been discussed in Section 5.1.2. For paper grammars, the guided path is the order of grammar entries chosen by the author. A dynamic approach does not have a guided path per se, as grammar entries are not ordered but merely belong to a domain and have a priority value. Nordhoff (2008, p. 316) presents a solution he used in his GALOES system. Each grammar entry is tagged with a question of the Lingua Descriptive Series Questionnaire. This way, the reader has the option of following the linear path of the questionnaire. This approach, however, may hold some redundancies as entries may contain answers to more than one question of the questionnaire.

Other scholars, such as Musgrave and Thieberger (2012, p. 64) and Drude (2012, p. 168), advocate for at least partial linearity in DRGs. Drude proposes a nested structure for grammar entries each represented by one hypertext page. Each page would have a list of links to their sub-pages and, in order to form a path, almost all pages (except the first, the last, and all top-level pages) have a link to the following, the preceding, and the super-page. Drude's and Nordhoff's dynamic path approach could in fact be merged, as a linear structure could be additionally attributed to a domain and receive a priority value. This way, the reader can be guided by a static path or explore the dynamic path.

Another value affected by the lack of linearity is what Nordhoff (2008, p. 305) calls "ease of exhaustive perception". A paper book can be read from the front cover to the back cover. Readers can easily determine where they are in the description and when they have read every page once they reach the back cover. This is similar for DRGs with a static path, as the last page can be reached at the end of the path. A dynamic structure does not necessarily provide a way for readers to signal how many pages they have read or when they have read the last page. Yet, some readers strive to read a complete grammar as they may be interested in the language family or in comparing the target language with other languages.

Ease of exhaustive perception WE VALUE THE AUTHOR'S DESIRE TO READ A GRAMMAR IN ITS FULL EXTENT.

Maxim 29 *A reader should know how much of the grammar they have already read and if they have finished. (Nordhoff, 2008, p. 305)*

Nordhoff (2008, p. 316) suggests a solution for DRGs with a dynamic path. All entry headings can be displayed on one page and formatted differently when they have been read. This way, a reader could see which and how many pages are left to read and if they have read all topics of a domain or of the entire grammar.

5.4 The structure of a DRG

Having discussed properties and functionalities of a DRG in the previous sections, this section will now assemble the three parts of a grammar and add the functionalities to sketch out an entire DRG. It aims at a more concrete exemplification of a DRG and the illustration of how the different parts of a DRG could interact. In some places, two or more possible constructions are presented in order to point out alternatives. While the GWT will not be developed within this project, the grammar format and a GET have been developed and will be discussed in detail in Chapters 6 and 7.

As mentioned in Section 4.5, a grammar includes three processes: writing, storing, and reading. Therefore, a DRG in three parts has been proposed: a GWT for writing, a grammar document in a grammar format for storing, and a GET for reading. The following subsections will discuss each of these three components in detail: Section 5.4.1 discusses the grammar format, Section 5.4.2 discusses the GET, and Section 5.4.3 discusses the GWT. The final subsection outlines a combined approach where the GET and the GWT are developed as one platform that authors can use to write grammars and readers, to explore multiple grammars.

5.4.1 The Grammar Format

The grammar format for DRGs is a standardized way to encode the grammatical description using a digital medium. All information captured in a grammar will be encoded in this grammar format in a document that has been referred to as “grammar document”. As a grammar format will be developed and discussed in Chapter 6, this section merely gives a brief overview of the structure of the grammar document.

One advantage of a digital format mentioned in Section 5.3.1 is the separation of the story and the database. This way, the structured information can be extracted much more easily by other applications or readers with a preference for the database structure. For this project, it means that the main grammar document containing the grammatical description is separated from the structured information that will be stored in databases. In order to grant consistency and easy interaction between the grammatical description and the databases, they can all be encoded using the same format[4]. What has been called the “grammar document” will thus be separated into seven documents:

1. The main grammar document containing the grammatical description and referencing entries of the databases

2. The terminology database containing terms and their definitions, hierarchically structured as one or multiple ontologies

3. The abbreviation database containing abbreviations and their expansions

[4]A DRG should also have a place to store multimedia resources (a multimedia database). However, these resources are not text resources. Therefore, they do not adhere to the grammar format and they are not included in the following discussion.

4. The bibliography database containing bibliographic entries referenced in the main document

5. The text database containing texts (ideally glossed and translated)

6. The example database containing example utterances taken from the texts and edited for didactic purposes

7. The lexical database containing lexical entries

These seven documents separate different types of information that are captured by a grammar. The center of these documents is the main grammar document. It contains the grammatical description with the network of grammar entries and references information saved in the databases (Figure 5.1). While the GWT's task is to fill the documents with content and to modify it, the GET's task is to extract the information of these documents and display one coherent grammatical description as encoded in the main grammar document. References to the databases will be resolved and the referenced information will be displayed within the grammatical description as known from paper grammars. Moreover, the GET can also provide direct access to the content of the databases.

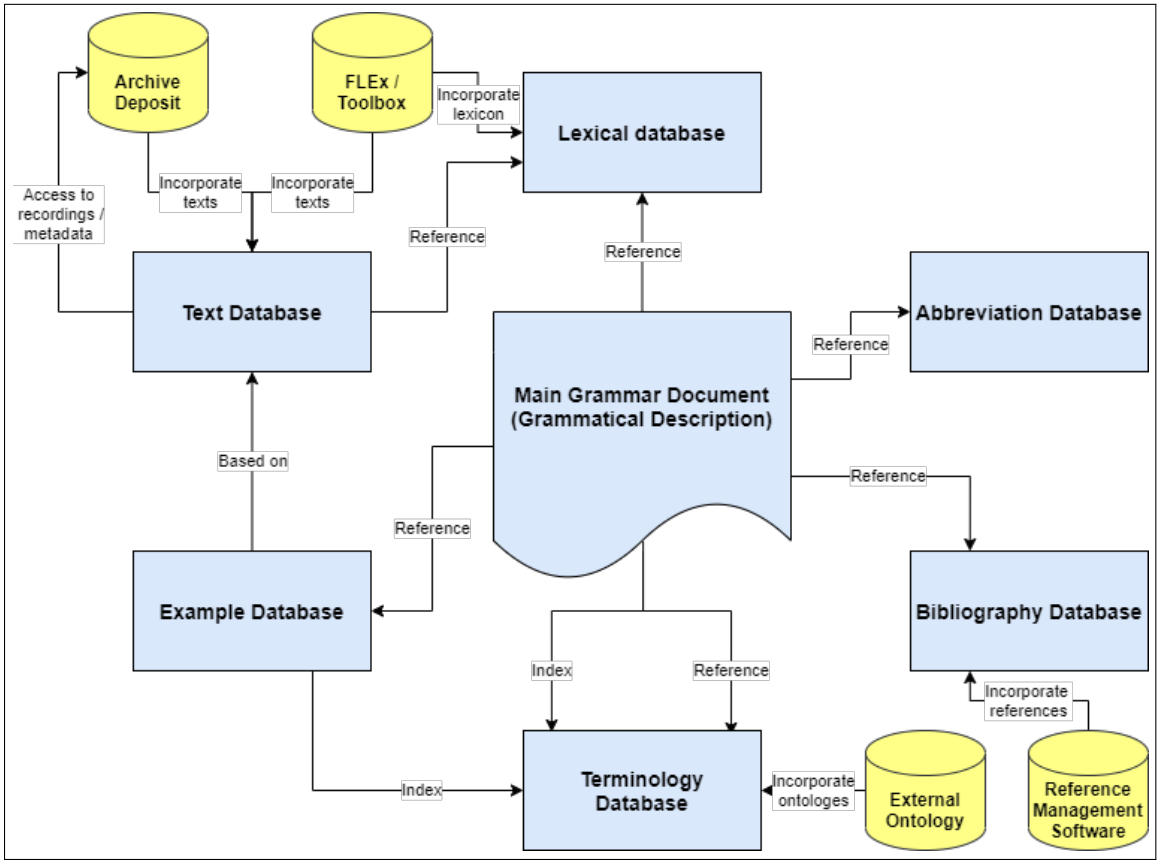


Figure 5.1: Interaction between the grammar documents (blue) and external resources (yellow)

5.4.2 The Grammar Exploration Tool

A GET is an application that reads the static grammar documents, displays them in a dynamic way and provides functionalities that improve the reading experience. For this model and the implementation discussed in Section 7, I will follow a hypertext approach, where the grammar's content is displayed on hypertext pages of a website. As the content of the grammar has been distributed between the main document and various databases, the GET should be able to display the grammatical description on the one hand and grant direct reading access to the databases, which could also benefit certain readers. In order to retrieve grammar entries, examples, or other data stored in the main document and the databases, a GET needs to be able to query through them. For this purpose a query language appropriate for the grammar format's markup language needs to be used.

In order to continue this discussion, we must first distinguish two types of GETs. The first option is that a digital publisher or institution, such as an archive, or a university develops a central grammar exploration platform that could be used to showcase multiple grammatical descriptions. The second option is a GET developed specifically for one grammatical description and tailored to its needs. The more efficient option is the development of a grammar exploration platform. This platform could be combined with a GWT and provide a way for linguists to write grammatical descriptions and publish them online. The linguists would not be concerned with the development of any application but could use both tools as a service. Some authors or language communities might still want to have the grammatical description on their own website – for instance, if the website combines different materials about the language and culture, such as the Interactive East Cree Reference Grammar (Junker, 2000).

A central grammar writing/exploring platform could additionally offer API services that enable external websites to access the grammatical descriptions and the displaying functionalities of the GET. This way, a grammatical description could be displayed on an external website without re-developing a GET. As the development of a GET is not as difficult as the development of a GWT, linguists with programming skills might want to develop their own GET that is tailored to their grammatical description. The linguist could use the central grammar writing/exploring platform to write the grammatical description in a standardized format. This way, the GET developed by the linguist could rely on the format and use the semantic markup for its own functionalities, as preferred by the linguist. Such a GET was developed in the scope of this PhD project in order to visualize the format and to exemplify the structure of a GET.

This section conceptualizes a GET that only displays one specific grammatical description. The combined approach for developing a central grammar writing/exploring platform will be discussed in Section 5.4.4, together with how that GET differs from the one conceptualized in this section. The following subsections will discuss how a GET could be structured as a website (Section 5.4.2.1), how grammar entries could be rendered (Section 5.4.2.2), how access to the databases could be granted (Section 5.4.2.3), and what kind of navigation tools could be provided (Section 5.4.2.4). While this section

should only convey an abstract impression on how a GET could function, all of these topics will be discussed in more detail in Chapter 7.

5.4.2.1 General website functionalities

A website usually has a home page where the reader is welcomed and introduced to the content and the function of the website. From the homepage, the reader can access other parts of the website. Besides the homepage, a GET website could have the following parts:

- A place where the grammatical description can be read (Section 5.4.2.2)
- A place or places where the different databases can be accessed (Section 5.4.2.3)
- Navigation tools that help the reader find information within the grammatical description or the databases (Section 5.4.2.4)
- A place where background information is given about the website and its content ("about pages")
- A registration and login facility where users could register and login if access is restricted

The GET's homepage could inform the readers of the most basic facts about the language and invite them to start reading by using one of the navigation tools. An additional button could have the title "start reading" and direct readers to the first grammar entry. Such a button is helpful for readers that intend to read the entire grammar or readers that want to familiarize themselves with the layout of the grammar entries before searching for specific topics.

The homepage could also direct to a few pages with background information about the website and its content. These pages will be referred to as "about pages". Readers might want to read about the language community, the language itself, the background of the grammar, the documentation project, or the responsible parties first before diving into the grammar's content. The content of these about pages is in fact identical to the background sections of a paper grammar (Section 6.3.1.1). Links to the about pages on the homepage could simply direct to the respective background section. Moreover, a help page is required to explain how the website can be used. It should thus inform the reader about all functionalities of the GET.

A registration and login functionality is required if access to the grammatical description is restricted. In that case, the user would have to create an account and ask for permission or, if the GET is commercialized, pay a fee to read it. User registration is also required if the application should save which pages the reader has already read.

Offline usability has been discussed in Section 5.2.2. The content could be saved in the browser and readers in remote areas only need internet access once to retrieve the data. After that, the browser can simply be used to read the grammar and look up information.

5.4.2.2 Grammar entries

The display of grammar entries requires a uniform layout that includes the content of the grammatical description, important metadata in a prominent spot, and navigation tools to navigate through the description.

As grammar entries have been established as the central unit of a grammar within this project (Section 4.3), it is also an appropriate unit to be rendered as one hypertext page. As a dynamic website, the GET could produce a separate hypertext page every time a new grammar entry is requested. This means, every time a reader clicks on a button to retrieve a certain grammar entry, the application produces a new hypertext page filled with the requested content. All of the elements within the grammar entry are displayed, and each time an element of the databases is referenced, its content is retrieved from the database and displayed within the hypertext page. For example, the reader can request the grammar entry about the focus particle (Figure 5.2). The GET searches for the re-

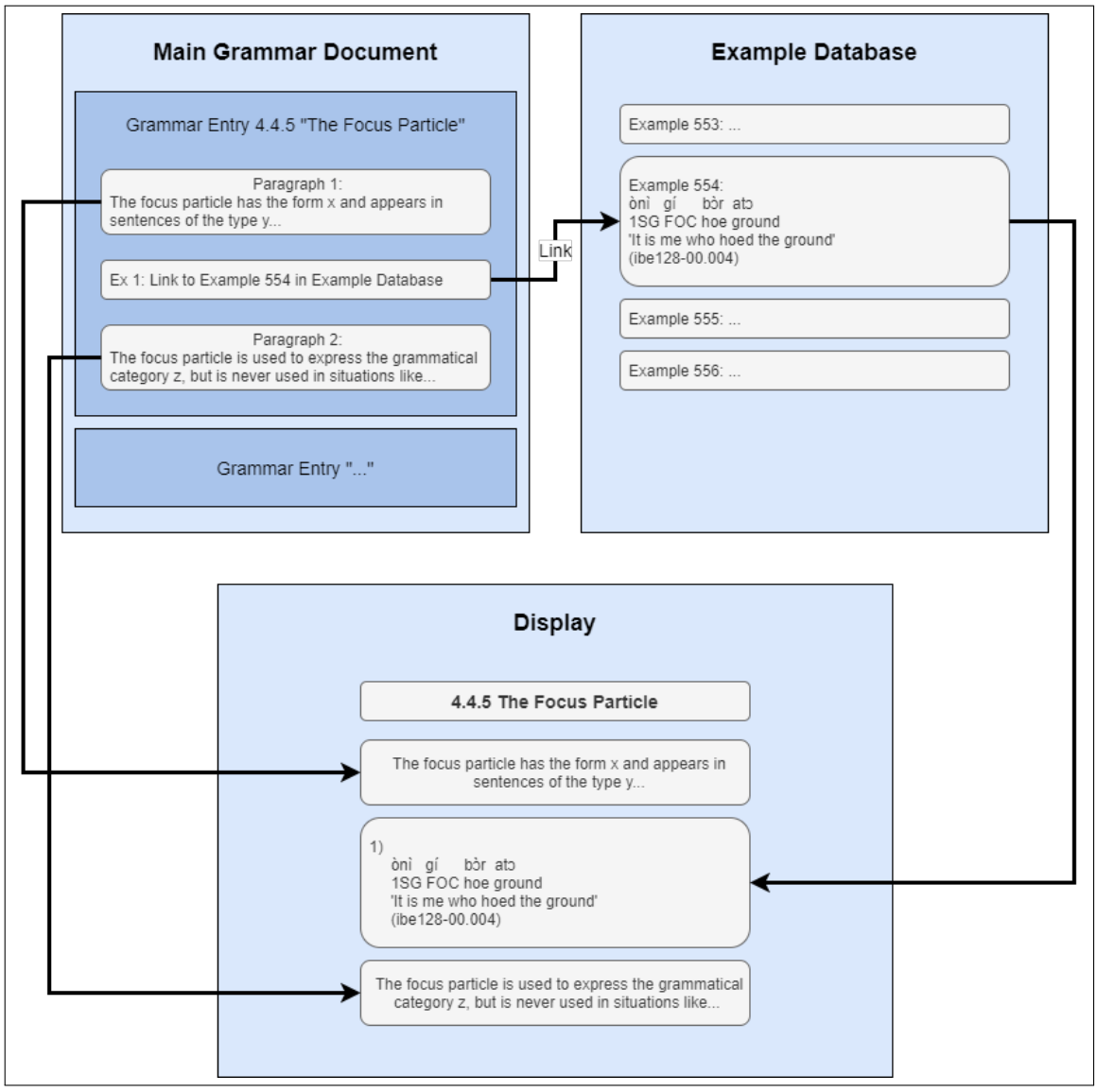


Figure 5.2: The display of a grammar entry

spective grammar entry within the main grammar document, retrieves all elements of the entry, and displays them on a hypertext page that was generated for this entry. The entry in question contains two paragraphs and an interlinear example. While the content of the two paragraphs can be retrieved from the main grammar document itself, the interlinear example is represented as a link to an example within the example database. Its content can be found in the example database and will be displayed after the paragraphs on the hypertext page.

Metadata is data about the grammar entry. If displayed at a prominent spot (for instance, the top of the page), the reader can use it to judge if the grammar entry contains the sought-after information and is worth being read. Metadata includes the heading and tags like the keywords used for the index, quality stages, or relative importance (Section 5.2.1).

General navigation tools, as described in Section 5.4.2.4, should be accessible from the grammar entry pages, but also internal navigation tools are required to navigate from grammar entry to grammar entry. These tools are buttons for cross-references, for the following and the previous entry, the super entry, and a list of buttons for subentries.

In accordance with the version control system discussed in Section 5.3.4, each grammar entry should have a place where edits of the page can be retraced and attributed to an author. This is especially important for co-authored grammars, as the reader needs to know which part has been contributed by which author. This could, for example, be implemented as a tab like in Wikipedia entries (Figure 5.3)[5].

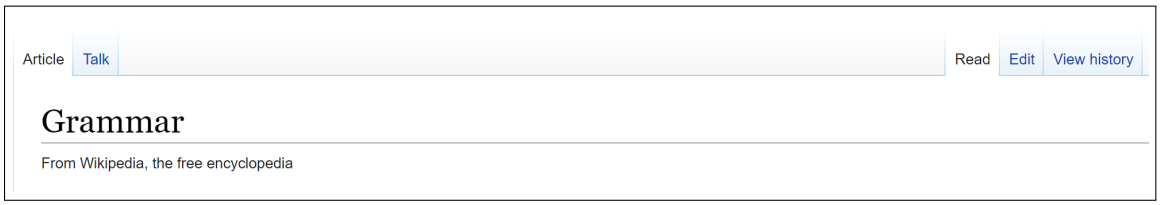


Figure 5.3: A Wikipedia entry including the version history tab in the top right corner

5.4.2.3 Access to the databases

As the grammar consists not only of the grammatical description but also of databases for texts, examples, lexical entries, terminology, abbreviations, and bibliography, these databases could also be made accessible to the reader. The layout for the display of each database may look different as it depends on the benefit a reader might have from accessing each database. As the access pages to each database is discussed and designed in Section 7.3.3, this section will merely exemplify the access page to the dictionary database to convey an impression of its purpose.

[5]The figure is an extract from the Wikipedia entry “Grammar” (retrieved 2020-12-10, from `https://en.wikipedia.org/wiki/Grammar`)

Words from the target language are often found within the text of the grammar itself, for instance, when a text is discussing the word's grammatical function or when the word used to exemplify a certain phenomenon. If these words in the main grammar document are linked to lexical entries in a lexical database, the GET could also create hyperlinks between words in a displayed grammar entry and the lexical database. This way, readers that want to look up further details about a word can click on the word and be guided to the full lexical entry with all related information. To achieve this, a separate place on the website needs to be created to access the lexical database. This might be a hypertext page that is created dynamically for each lexical entry. Figure 5.4 shows the layout for the lexical database access page in the exemplary GET for Abesabesi. In this case, the entry *ìkpàr* 'children' has been selected. The reader can call the access page directly from the homepage and select an entry from the list or through the search field. Moreover, if a word in the target language is clicked in the grammatical description, the GET directly guides the reader to this access page with the respective lexical entry already selected.

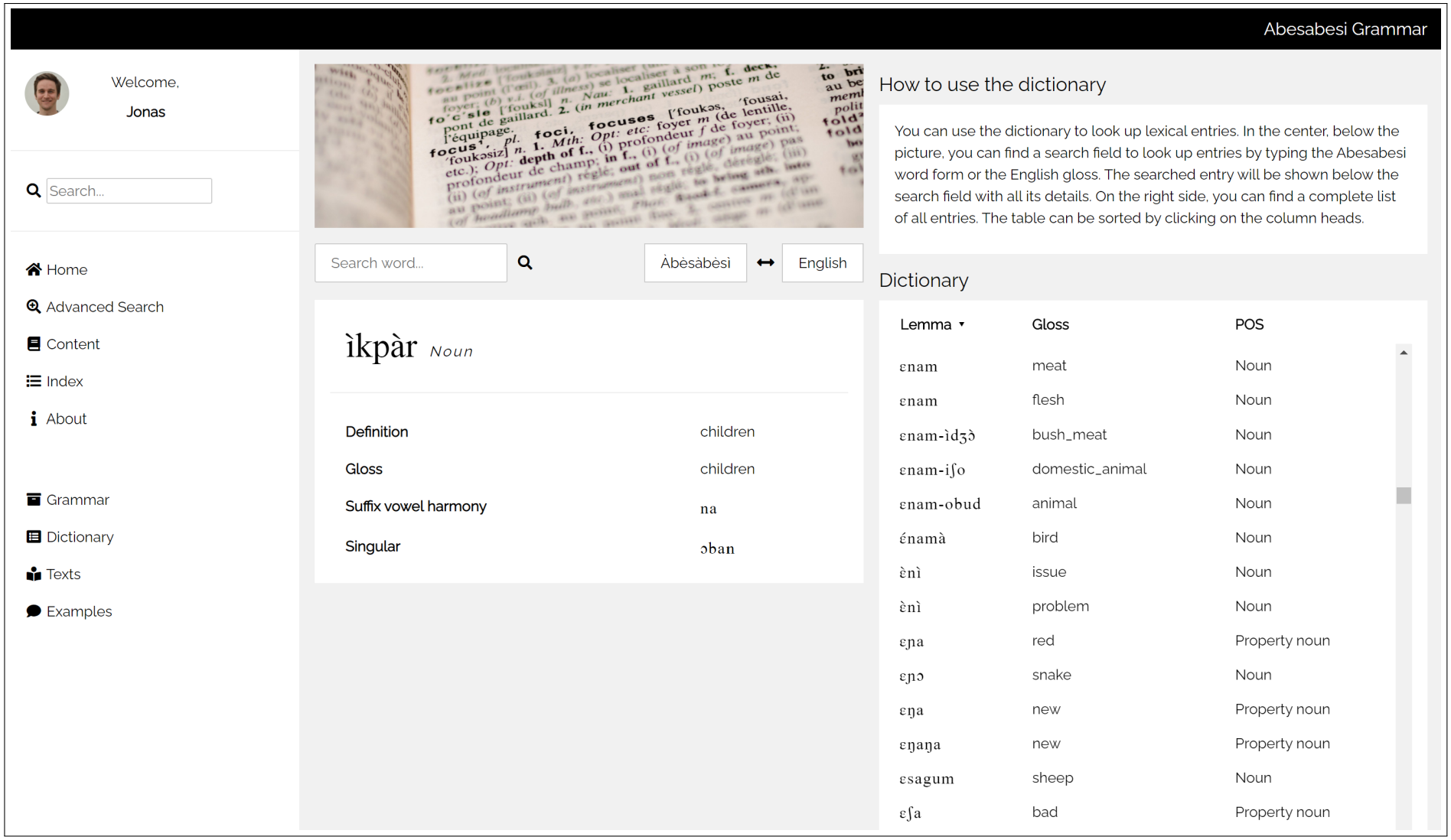


Figure 5.4: Dictionary layout

5.4.2.4 Navigation tools

The main task of a GET, besides displaying the grammar entries, is to help the reader find information within the grammatical description. For this task, a GET can be equipped with traditional navigation tools, such as a TOC and an index, as well as a search tool, version history, and layer selection.

A TOC could be designed in a similar way to a traditional book TOC with a nested, numbered list of the headings. Additionally, headings could direct the reader to particular grammar entries. Deeper levels of the TOC could be hidden for better clarity until a reader unfolds them. Figure 5.5 displays the TOC of the Abesabesi Sketch Grammar. On the left side, a TOC with only the first-level headings is shown. If the triangle to the right side of a heading is clicked, its sub-headings are displayed (TOC at the right side).

If a strict separation of form-entries and function-entries is undertaken, as proposed by Nordhoff (2012b) (Section 6.1.3), this could be reflected in the TOC. Each form-entry could be clicked to display the related function-entries and vice-versa. Figure 5.6 displays the French TOC used in Nordhoff's example. As the grammar entry "Inversion-construction" is a form entry with corresponding function entries, it has a double arrow symbol that can be clicked. When clicked, a modal window that shows all corresponding entries appears.

Nordhoff's (2008) dynamic approach could also be reflected in the TOC. The domains could be presented in a list with all related articles subordinated in the order of their priority value. Additional information given in the TOC could include metadata tags and indications for whether a page has already been read or not. This is especially important for the dynamic approach, which does not provide a static path. Read pages in the TOC could be marked by formatting them differently. Similar to clicked links, the text color could change.

The index lists all terms used within grammar entries or as keywords for grammar entries or examples. Like in paper grammars, a number could indicate the grammar entry where the keyword is used. In a DRG, these numbers could also direct the reader to the respective grammar entry. If the example database is also tagged with keywords, links from the terms to their respective examples could be provided. The definitions of the terms could be given when clicked on, either as a modal window or as a separate page.

A GET could provide different search functionalities to guarantee fast access to the desired information. A simple full text search form that is accessible from every page of the website could satisfy the needs of readers that value a fast process. A word can be typed into the form and, by submitting the query, all occurrences of that word could be listed. The results could also be prioritized by context: An entry with the searched word in the heading might have top priority, followed by an entry where the word appears frequently, etc. This simple search form could be separated from an advanced search option. Users that require very specific information and value an exact searching method could use the advanced search. Contextualized and combined queries can be used, such as "find a grammar entry that is tagged with the expression 'vowel harmony' and that

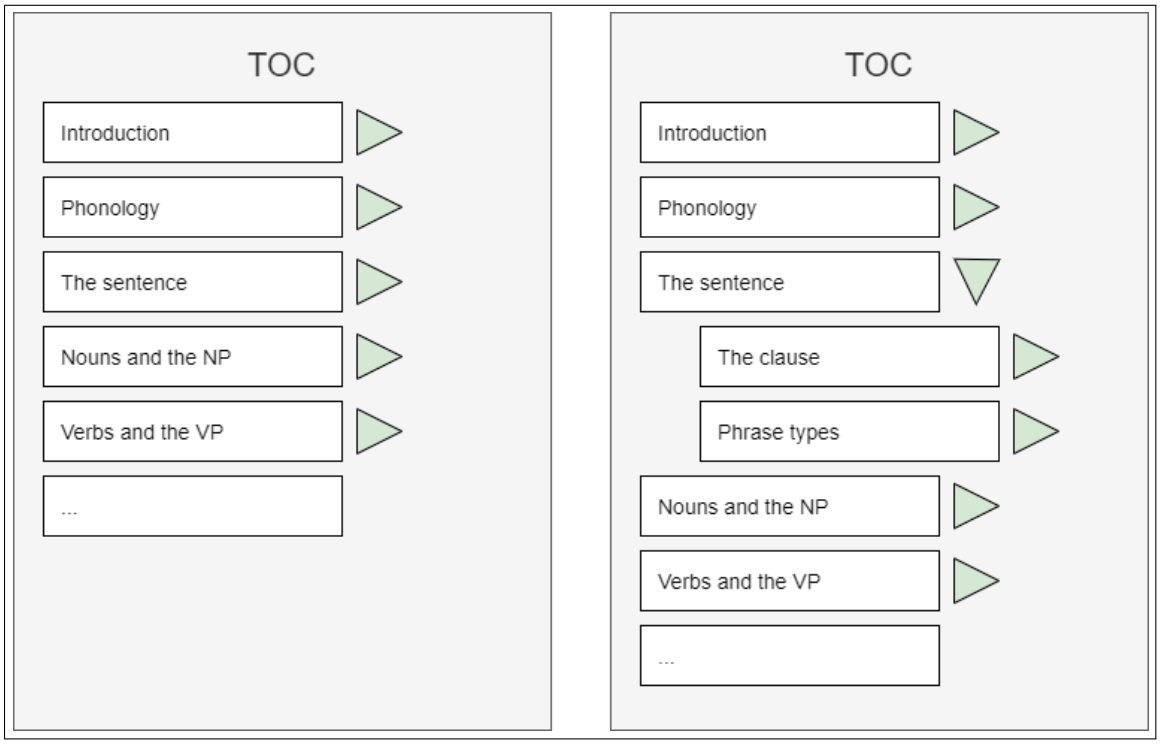


Figure 5.5: TOC - Unfoldable subheadings

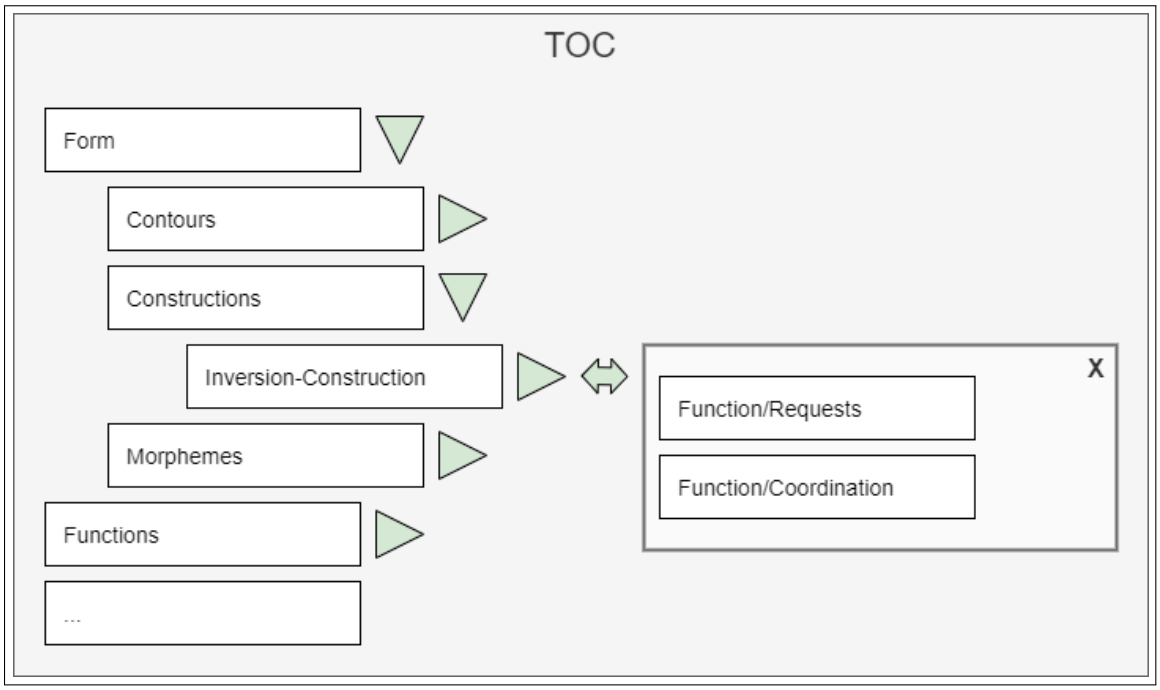


Figure 5.6: TOC - form-function relations

contains an example tagged with the expression 'perfective'", or "find a grammar entry that contains the expressions 'focus' and 'negation'". Moreover, the option to use regular expressions could confine the search results.

Version control and version history have been explained in Section 5.3.4. In order to access older versions of the grammatical description, the GET needs to have a version history. This could be a separate page where all older versions and the current version are listed and can be clicked to be directed to an older version. In order to remind the readers in which version they are and to facilitate citing, each grammar entry page should display the version number that is displayed.

If the grammar is multi-layered (see Section 5.2.8), the GET needs to have a way to select a specific layer. For instance, if grammar entries are available for different dialects of the language, the reader needs a way to select the desired dialect. This could be done with a dropdown menu. After selecting the dialect, all grammar entries will relate to the selected dialect. An additional functionality would be parallel views to compare a phenomenon in two or more varieties.

5.4.3 The Grammar Writing Tool

A GWT needs to assist a grammar author in two tasks: writing a grammar and managing the related databases. It needs to contain a grammar editor to write and edit the grammatical description and several tools to manage the different databases.

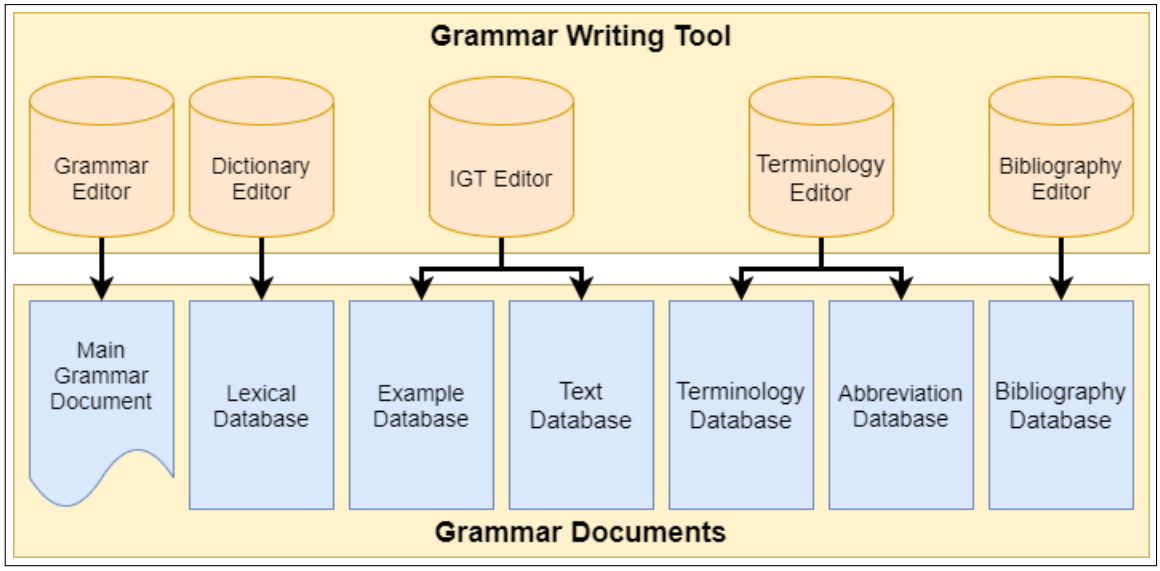


Figure 5.7: All parts of the GWT and their interaction with the grammar documents

5.4.3.1 Grammar editor

The main goal of a grammar editor is to guide authors to produce grammatical descriptions that can be stored in a standardized format. While the authors may find complete

expression in terms of the grammar's content, they need to work with standardized but customizable grammar bricks in order to allow a GET to display the description and other applications to exploit the grammar's data. The main features of a grammar editor are:

- to provide structural bricks that can be filled with content (heading, paragraphs, examples, etc.)
- not to interfere with the grammar's content
- not to bother the author with the grammar format
- not to bother the author with layout decisions
- to remind but not to force the author to use best practices of grammar writing

Examples for comfortable editors that do not bother the author with the format itself are the XMLmind Editor that can be extended with XLingPapers (Section 6.1.4) or the Oxygen XML Editor, which is popular for editing TEI documents. Both editors provide a way to work on XML documents without being confronted with the format itself. Instead, the author sees a formatted version of the document. This method provides the familiarity of a WYSIWYG editor, but forces the author to write the grammar in a format-conforming way and to use the structural elements provided by the format. Authors familiar with mark-up languages might chose a view that lets them directly edit the XML format.

Following Maxim 17, the editor should not bother the author with the layout options. However, this does not mean they should be prohibited from making layout decisions. Authors might want to chose a specific layout for an element in order to follow didactic goals. This needs to be taken into account. The editor of the GWT should therefore provide a default layout for the structural elements but also allow the author to change the layout or parts of it.

The main goal of the editor is to guide the author in writing the grammar in a format-compliant way. It thus provides structural elements. In order to allow for a grammatical description that follows the language's own "genius" (Sapir, 1921, p. 127) and that does not impede the author's creativity, the editor should not interfere with the grammar's content. It may propose outline templates and content suggestions to inspire the author or remind them of topics they might have forgotten, but should never force the author to add a specific kind of content. XLingPapers, for instance, provides a grammar outline by Payne (similar to the one Payne (2014) provides in his appendix). When creating a new document, the author can select the outline as a document type and the new document will already contain the sections and headings of the outline. The author can still change these headings or even the whole structure, but can be guided by the outline. Similarly, templates can be provided for smaller units such as for examples, tables, or linguistic units. Reminding the author of adding optional elements, such as metadata tags, increases the chances that the author will improve the grammar by adding information about each grammar entry.

Layering of the grammar also needs to be enabled by the editor. At first, the author needs to add layers to the grammar. For instance, they could add Russian as another metalanguage to the grammatical description. Once elements are added to the grammar, the editor needs to provide a way to add another version of the element with Russian content.

As several databases are separated from the grammatical description (Section 5.4.1), the GWT needs tools to manage the databases. These tools are presented in the following sections. All databases managed though theses database tools need to be accessible via the grammar editor, as authors need to reference the entries of all databases. For instance, while editing a grammar entry, the author might want to cite an article. After adding a reference element, the editor needs to open a search field where all references from the bibliography database can be accessed. After clicking on the desired bibliographical entry, the reference will link to the entry. Once the grammar is read in the GET, the citation is displayed in a standardized format and can be clicked to view the full bibliographical entry.

5.4.3.2 Terminology and abbreviation management

A tool to manage abbreviations and terms could be combined. Authors should be able to add terms, their abbreviations, and/or definitions. Terms can be ordered hierarchically and separated into different ontologies as discussed in Section 6.3.3.2. Moreover, the author should be able to import existing ontologies, such as GOLD (Farrar & Langendoen, 2003), OLiA (Chiarcos & Sukhareva, 2015), or OnLiT (Lehmann, Chiarcos, & Hellmann, 2017) to use their terms in the grammatical description.

5.4.3.3 Bibliography and dictionary management

A bibliography management tool within the GWT can be used by the author to import bibliography collections from other applications the author might use or to manually enter bibliographic entries.

The dictionary tool can be used to manage lexical entries. It needs to be able to import lexical databases from other applications such as FLEx or Toolbox. Additionally, it should allow authors to enter lexical entries manually or edit existing ones.

As bibliography tools (Zotero[6], Mendeley[7], etc.) and dictionary tools (FLEx[8], Toolbox[9], etc.) already exist and function similarly to how they would function in the GWT, they are not discussed in detail.

[6]retrieved 2020-12-11, from`https://www.zotero.org`
[7]retrieved 2020-12-11, from`https://www.mendeley.com`
[8]retrieved 2020-12-11, from`https://software.sil.org/fieldworks/`
[9]retrieved 2020-12-11, from`https://software.sil.org/toolbox/`

5.4.3.4 Example and text management

As the example database and the text database both contain Interlinear Glossed Text (IGT) (explained in Section 6.3.2.2), they can be managed by one editor whose main function is to process IGT. Functional requirements for a general IGT editor have already been discussed and mapped with existing approaches in Hughes, Bow, and Bird (2004). The editor described in this thesis has additional requirements to accommodate for tasks such as the management of two databases. This section presents a very simplified workflow of this IGT editor (illustrated in Figure 5.8). A more detailed account on processing IGT will be given in Section 6.5.5. The IGT has the following tasks:

1. Import transcribed utterances from an archive deposit/from transcription software
2. Create links between the imported utterances and the recording and metadata in the deposit
3. Process IGT
    - Let authors modify and format the transcribed utterances
    - Let authors manage the utterances by organizing them as texts in the text database or as examples in the example database.

Transcriptions and translations usually already exist when writing a grammar. Instead of writing interlinear examples from scratch, an editor could simply import transcribed utterances from a deposit archive, where transcriptions are often stored along with recordings and metadata. An IGT needs to be able to access an archive deposit in order to import the transcriptions and to connect the IGT with recordings and their metadata in the archive. Moreover, an IGT editor should be able to import utterances from ELAN files and link them with the time-aligned snippet of the recording. An author could access ELAN files through the IGT editor, select a segment, and import the text of selected tiers from the file into the tool. Once the tiers are imported, the example must be linked to the corresponding time frame of the recording. This way, a reader of the grammar can read an example in the GET while the correct snippet of the recording is played.

Other required functionalities are the manual input of examples and the extraction of examples from the text database. If no recording exists, the author should be able to add an example manually. An IGT editor should also enable authors to create new IGT. The production of text collections could work as described for examples. The author would only have to decide whether an imported or created utterance should be part of a text and be stored in the text database or if it should be used as an example and stored in the example database. In order to process and store IGT, a specific format is required. Existing IGT formats will be discussed in Section 6.5.5 before proposing a new IGT format that is required for the needs of this project.

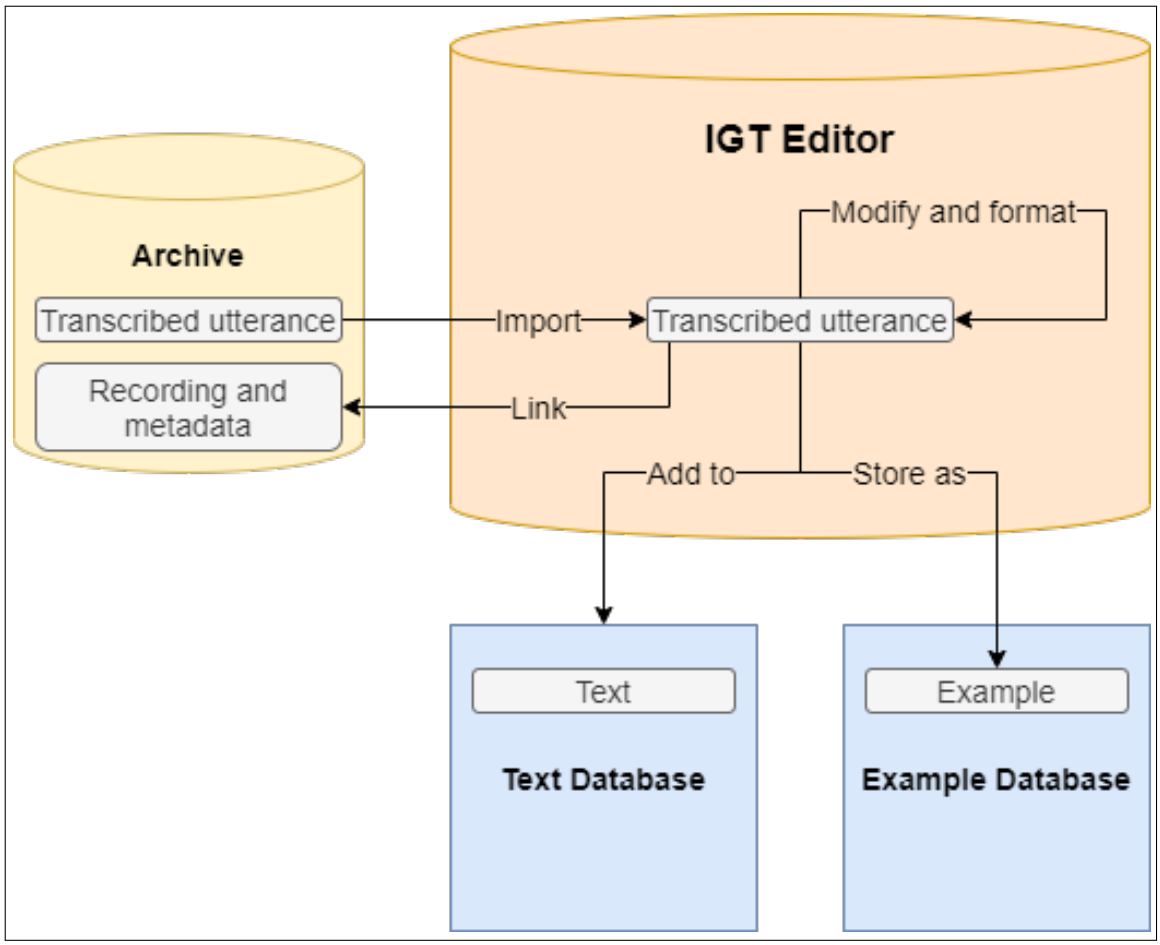


Figure 5.8: Workflow of the IGT editor

5.4.3.5 General functionalities

Besides the editor and the database tools, a GWT needs to offer functionalities that enable an efficient workflow. As the DRG is supposed to provide multiple versions to assure consistent citing, the author needs to be able to publish edits as separate versions. A version should encompass not only the current state of the grammatical description, but also that of all databases. Before publishing a number of edits as its own version, the author needs to have the possibility to save all edits, especially when working offline. Even if edits are not saved, the GWT should produce regular automatic backups in order to prevent data loss (see Maxim 24). Backups could be saved locally when offline and on a server when online.

Distributed authorship, as discussed in Sections 5.3.2, is only possible if the GWT allows more than one user to edit the grammar. Therefore, it needs to have a versioning system. Different authors need to be able to retrieve the current version of the grammar, edit it, and upload the updated version. If different authors work on the grammar at the same time, the GWT needs to provide a way to merge the two versions when uploaded. If no conflicting entries are detected, this can be done automatically, but if, for instance,

two authors have edited the same grammar entry, the merging process needs to be done manually. The GWT should also guide the author that uploaded a conflicting version through the merging process.

5.4.4 Combined approach

This section proposes a combined approach, where one centrally developed platform can be used for multiple projects. The platform combines a GWT and a GET, which can be accessed from the same website. This combined platform will be referred to as "grammar platform". Developing a GWT only makes sense if it is usable for multiple projects, as the development is rather complex and it would become obsolete once the grammatical description has been completed (unless it is extended or updated). This is why the functionalities for the GWT are the same as proposed in Section 5.4.3. Section 5.4.2, however, has concentrated on the development of a GET tailored to one specific grammatical description. Thus, the functionalities of the GET designed for the display of multiple grammars differs slightly.

A homepage could let the user choose whether they want to be directed to the GWT or the GET. A login area could allow the users to be identified and be automatically directed to where they have left off – be it in the GET or in the GWT. User management is a requirement for a grammar platform, as grammar writing projects have to be attributed to specific authors. A registered user could start writing a grammar and appoint other users to be co-authors. Registering would not be mandatory for using the GET, unless the grammar platform is commercialized or a certain grammatical description has access regulations. A benefit of registering before reading a grammatical description is that the application can memorize which pages have been read and the user can be directed to the part where they have left off the last time they signed in.

As mentioned before, a GET that displays more than one grammatical description requires a few different functionalities than one that is tailored to one particular grammatical description. First, it needs a starting page where all available grammatical descriptions are listed. The platform can offer different ways to browse through the list of grammatical descriptions. They can be ordered by region or by language family and a search field could help find the sought after grammatical description. After selecting a grammatical description to read, the user should decide which version they want to read. Now, the user could be directed to a homepage, like the one described in Section 5.4.2. A possible difference could be the (advanced) search functionalities, where queries could be extended over multiple grammatical descriptions. This would enable typological research across languages. A grammar platform as described in this section could thus be counted as a cross-linguistic approach as discussed in Section 4.6.5.

The combined approach resulting in a centrally developed and maintained grammar platform is probably the most realistic vision for DRGs. This is because the development of a GWT is a rather complex task and would require several developers, possibly the cooperation of multiple institutions, and a substantial amount of funding (Drude, 2012,

p. 161). A project of this size could only be funded if a long-term infrastructure like a grammar platform would result from it. It would need to be directed by at least one institution to ensure stability and longterm maintenance. The resulting grammar platform would be a great contribution to enhance linguistic research. Grammar writing would be facilitated, grammatical descriptions could be collected in one place and explored with efficient new functionalities.

The goal of the next chapter is to propose a standard for encoding reference grammars that could potentially be used in such a grammar platform. The following Chapter 7 illustrates how this format could be displayed by presenting the development of a standalone GET application as described in Section 5.4.2.

Chapter 6

Towards a format for digital reference grammars

Having conceptualized a DRG and separated it into three main components – the grammar documents, the GET, and the GWT – this chapter now embarks on the development of a grammar format to encode grammatical descriptions into the digital grammar documents. Developing the grammar format is the main focus of this thesis, as it could be beneficial for future researchers. Although there has been an academic discourse about DRGs and how to develop a storage format for them, the few existing approaches were not elaborated on or are not widely used. The expansion of a format is mostly impeded by the lack of stable software to write and display the stored grammatical description. This is probably the reason why reference grammars are mostly written with a common text editor and published as PDFs. Nonetheless, there is an urgency for a digital storage format, as there are many benefits for storing reference grammars in a uniform manner and from distinguishing and linking discrete elements of a grammatical description. This way, the stored information can be made available for other applications and digitally enhanced research. This is why I will propose a format for DRGs and simultaneously emphasize the need for a large scale development project for a DRG infrastructure.

The proposed format follows an integrated approach and builds on existing infrastructure for text encoding – the TEI. It only uses existing elements of the TEI framework and proposes a method for how to encode grammars with it. This will be referred to as "Text Encoding Initiative Reference Grammar Format (TEI-Grammar)". Grammar writing has not been a focus in TEI although the need for a separate treatment of a grammar's idiosyncrasies have been noted (Drude, 2012, p. 172). This proposal will be the first comprehensive approach to the encoding of reference grammars in TEI. As with other TEI documents, a grammatical description can theoretically be encoded by the authors themselves. In the ideal case where a GWT has been developed, the encoding would only be done through the GWT without confronting the authors with the format itself.

After evaluating previous approaches to the development of a format and presenting the utilized technologies in Section 6.1, this chapter's methodology will be presented

in Section 6.2. Section 6.3 will model the structure of reference grammars in order to have a base for a grammar format. While developing the grammar format, parts of the Abesabesi Sketch Grammar have been encoded to exemplify and refine the format. This encoding process will be presented in Section 6.4 before proposing the grammar format TEI-Grammar in Section 6.5. The format's interaction with an application will be tested and evaluated in Chapter 7 where the development of an GET will be illustrated.

6.1 Previous formats

Recent discourse about digital grammaticography has produced a few suggestions on how a format for DRGs could be structured. The content of a hypertext grammar could theoretically be directly written into the HTML code, but there are several reasons to structure the grammar in a separate format. Although HTML is meant to contain the content of a website and is separated from its style, the language is still structured in regards to the presentation. It divides the document into structural website elements: headings, paragraphs, tables, and other elements. A general consensus is to avoid such a presentational format to store the grammatical description and to use a format where the content can be semantically structured. It is the task of a GET to present the grammar on a website and to convert the semantically structured format into HTML (Drude, 2012, p. 171).

A format that focuses on the semantic structure of a text could represent the text in a more fine-grained way and, additionally, could differentiate the linguistic units used within the text. Lexemes, for example, could be differentiated from word forms and grammatical categories could be annotated on morphemes, associated with paradigm tables, or with sections. This allows for detailed search queries in the GET and external applications could exploit the data for other purposes. As Good (2004) and Nordhoff (2012b) argue that grammatical descriptions are semi-structured texts, the information within a grammar is not only structured into textual elements such as headings and paragraphs, but also into tables, examples, word-meaning pairs, etc. Those structural elements require a more fine-grained representation than that of HTML.

Most format suggestions center around the use of a format based on the Extensible Markup Language (XML) (e.g. Good, 2004; Drude, 2012; Nordhoff, 2012b). After presenting the XML format in Section 6.1.1, Section 6.1.2, 6.1.3, and 6.1.4 will discuss three XML based formats. Good (2004) and Nordhoff (2012b) are two approaches where a theoretical model for reference grammars was formalized in a schema document. The modeling of a grammar's formal elements in Section 6.3 has been based on Good's model. An XML based markup language specifically for text encoding is the Text Encoding Initiative (TEI). After presenting this standard in Section 6.1.5, Section 6.1.6 will discuss TEI based approaches for encoding reference grammars. Finally, Section 6.1.7 will discuss an approach by Good (2012) using semantic web technologies and Section 6.1.8 will discuss a more resent approach using the Cross-Linguistic Data Formats (CLDF).

6.1.1 XML

XML is a standard published by the World Wide Web Consortium (W3C) used to store and exchange textual data. Within an XML document, data is stored in so called "elements" which constitute the main structural units of the document. Elements begin with a start-tag (e.g., `<section>`) and end with an end-tag (e.g., `</section>`) or are represented by an empty-element tag (e.g., `<section/>`). The tag name (like "section" in the previous examples) can be chosen by the user and is used to label the data that is stored within the element. Additionally, elements can have attributes, which are key-value pairs indicated in the start-tag of an element. Attributes are used to store additional simple information of the element. Each attribute can only be used once within a tag and can have only one value. A section element could for example have an attribute that contains an identifier: `<section id="1.5.3">`.

Elements can be nested within other elements to create a tree structure. XML documents thus also map hierarchical relations. A small example of an XML extract is given in Listing 6.1. It shows a chapter element with a nested heading, paragraph, and section, which also has a nested heading and paragraph.

Listing 6.1: XML example

```
<chapter id="1">
        <heading>Phonology</heading>
        <p>Lorem ipsum dolor sit amet, consectetur...</p>
        <section id="1.1">
                <heading>Vowels</heading>
                <p>...</p>
        </section>
</chapter>
```

Throughout this thesis, and especially in this chapter, XML and TEI element names will be indicated in angle brackets similar to start tags: e.g., `<div>`. Attributes will be indicated with an initial at sign (@): e.g., `@id`.

As there is not a fixed set of tags, XML is a flexible language that can be tailored to the needs of a specific community of users. However, a controlled vocabulary of tags can be standardized by using a schema. Parsers can validate an XML document to guarantee that a document is compliant with a schema. This way, XML based languages are machine-readable and human-readable at the same time.

XML has the advantage to be a widely used format language that has the best chance to stay in use for a long time, as it can adapt to many data types (Drude, 2012, p. 171). Most programming languages have an XML integration. A whole suite of XML technologies facilitate the use and processing of XML documents. At least two of them are essential to this undertaking and will be presented in the following paragraphs.

The Extensible Stylesheet Language (XSL) comprises three components: XSL Formatting Object, Extensible Stylesheet Language Transformations (XSLT), and XML Path Lan-

guage. Important for this project will be XSLT, which enables the transformation of XML documents. Using XSLT, an XML document can be parsed and rearranged or transformed into, for example, an Extensible Hypertext Markup Language (XHTML) document, which can be rendered by a browser. Instead of developing an application to read the grammar documents, a transformation sheet could be used to transform the grammar into hypertext pages. This transformation approach would be much easier but could lack many of the proposed features of a GET. This project therefore follows the approach of developing an application that operates with XML Query (XQuery). However, a transformation of the grammar format into XHTML can be useful for this project in that it could produce a linear formatted version of the grammatical description as a data output (Section 7.4.4).

The other XML technology important to this project is XQuery. It is a query language intended to search through large XML databases. It provides several query mechanisms such as loops, sorting, or the use of variables in order to find and restructure data from XML documents. Although it has roots in a different tradition, its functionalities overlap with XSLT. Within this project, it will be used within the GET application to retrieve information from the grammar documents and restructure it (Chapter 7).

6.1.2 Good (2004)

Good (2004) surveys four printed reference grammars and presents a general model of a traditional reference grammar's structure. The model is presented as an XML format. An exemplary XML fragment (see Listing A.1 in the Appendix) as well as a Document Type Definition (DTD) schema document are provided. As Good investigates a general structure for and common features of existing paper grammars, his model has been taken as a basis for modeling reference grammars in this thesis (Section 6.3).

Good's central unit within the grammar is what he calls "annotation". An annotation is synonymous with the term "grammar entry" used in this thesis and contains grammar bricks such as descriptive prose, exemplar data, etc. Annotations can be nested, resulting in the tree structure typical to reference grammars. Besides annotations, a grammar also contains a text database, a lexical database, and several ontologies that can all be referenced within the annotations.

Regarding elements within annotations, he distinguishes descriptive prose, exemplar data (separated into textual and lexical exemplars), and structured description. Descriptive prose can contain references to lexical items, other sections, terms drawn from ontologies, and exemplar data. Figure 6.1 shows a diagram of the model displayed in Good (2004).

As this approach aims at developing a format for reference grammars that could be accepted by a wide range of linguists, the proposed format needs to have a familiar structure and encompass all elements of a traditional grammar. Good's model will thus be taken as a base that can be extended with optional elements specific to digital grammars.

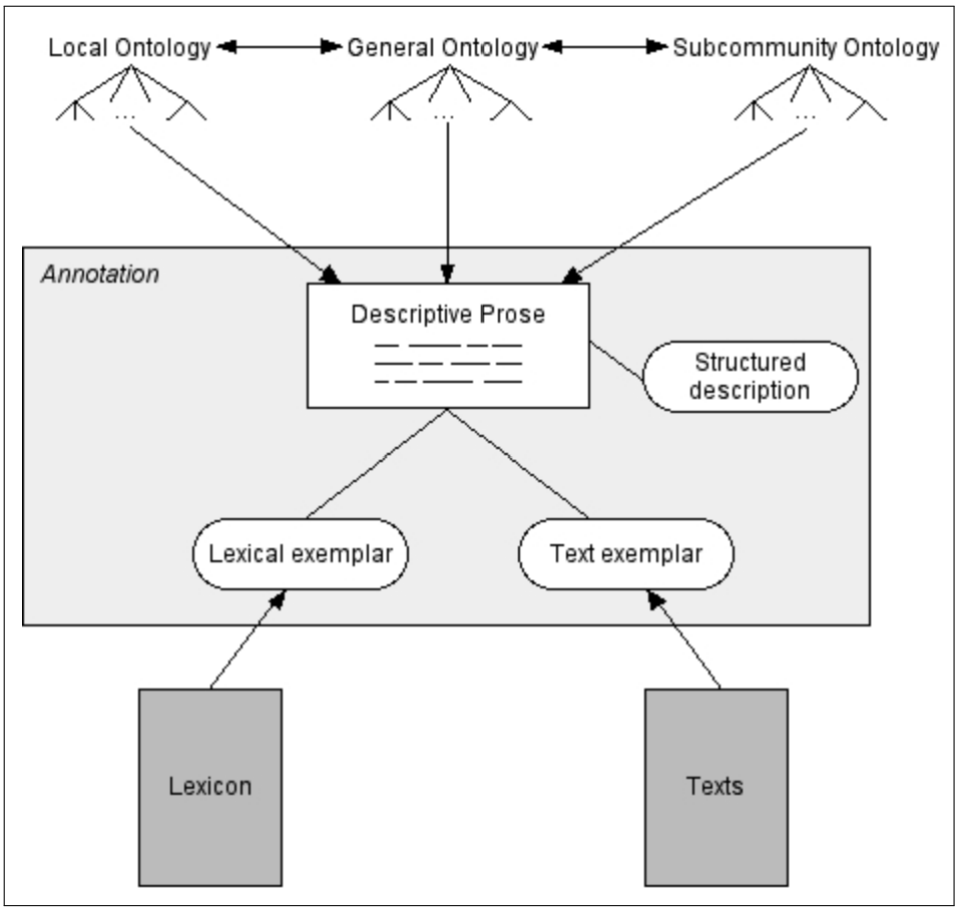


Figure 6.1: Good's (2004) model of reference grammars

6.1.3 Nordhoff (2012)

Nordhoff (2012b) uses a sample of eight reference grammars to propose his model of reference grammars. The model goes beyond the traditional structure of paper grammars as it formalizes the grammar as a collection of form-meaning pairs (FOMPs). Nordhoff bases his model on Good (2004). He also uses the annotation as a central unit but concentrates on a special kind of annotation, the FOMP. Based on the assumption that most morphosyntactic phenomena can be separated into a form and a function, a FOMP is the description of a phenomenon where the formal description is separated from the functional description. These FOMPs can either be directed from form to function (semasiologic approach) or from function to form (onomasiologic approach)[1]. Form-to-function FOMPs can be descriptions of morphemes, constructions, or intonation and are grouped into the collections "morphemicon", "contructicon", and "contouricon". Function-to-form FOMPs contain the description of semantic, pragmatic, or discourse-related phenomena that are grouped into the "semanticon", "pragmaticon", and "discoursicon" (see

[1]The differences between the semasiologic and the onomasiologic approach in grammaticography were discussed in Section 4.3.3

Figure 6.2). While form-to-function FOMPs have a formal unit as their lemma, function-to-form FOMPs have a functional unit as their unit. As form and function naturally have many-to-many relations, form-to-function FOMPs can be related to multiple function-to-form FOMPs and vice versa. This model constitutes a combined approach (semasiologic and onomasiologic), as described in Section 4.3.3.4. Through these mapped relations, readers can take full advantage of the semasiologic and onomasiologic approach.

As this model relies on a rather strict separation of form and function, this entails problems mentioned by Lehmann (2004a, Section 3)(see Section 4.3.3.3): Some categories are defined by formal and functional parameters. Where should these categories be discussed? This especially applies to the description of word classes. Word classes are usually defined by formal and functional criteria. If they are treated as part of a formal collection, with which functional units do they have relations? In general, it is rather time-consuming to structure a grammar that relies on clear relations between form and function where form and function are separated. How about irregularities or restrictions on when a relationship between a form and a function can be assumed? The same holds for phenomena that can only be described incompletely, as the amount of data or research does not allow for more. Relations to functional units can only be assumed for them. Even if this model tries to map important relations to make them machine-readable, it seems to be hard to implement. With the separation of form and function and the mapping of relationships, it is a rather restrictive formal approach to model reference grammars. However, if FOMPs are only special instantiations of annotations, they could be optional and treated like an extension to the proposed format.

6.1.4 Summer Institute of Linguistics

The Summer Institute of Linguistics (SIL) has developed a format specifically for the creation of linguistic texts. XLingPaper (H. A. Black, 2009) is an XML based format defined by a schema and an adaption to the XMLmind XML editor, which allows authors to edit XML files without being confronted with its syntax. Templates for textual elements and those elements that are specific to linguistic texts are provided to guide the author in writing texts and simultaneously store it in the underlying format. This way, authors do not require knowledge of the format's syntax. According to C. A. Black and Black (2012), XLingPaper facilitates the editing of linguistic examples, their numbering, citing, provides abbreviation management, and allows for common output formats such as PDF and HTML. These functionalities also facilitate grammar writing, which makes XLingPaper a good example for a GWT. The underlying format is a potential candidate for the grammar format, as it contains the majority of the formal units of grammars. It also includes elements for linguistic examples and language data. Semantic markup for the differentiation of linguistic units, however, is not included. The author can merely mark text as language data but cannot differentiate morphemes from lexical entries or word forms. Moreover, tagging a section for metadata is only possible for indexes. Other key-value pairs, such as quality assessment and relative importance cannot be attributed

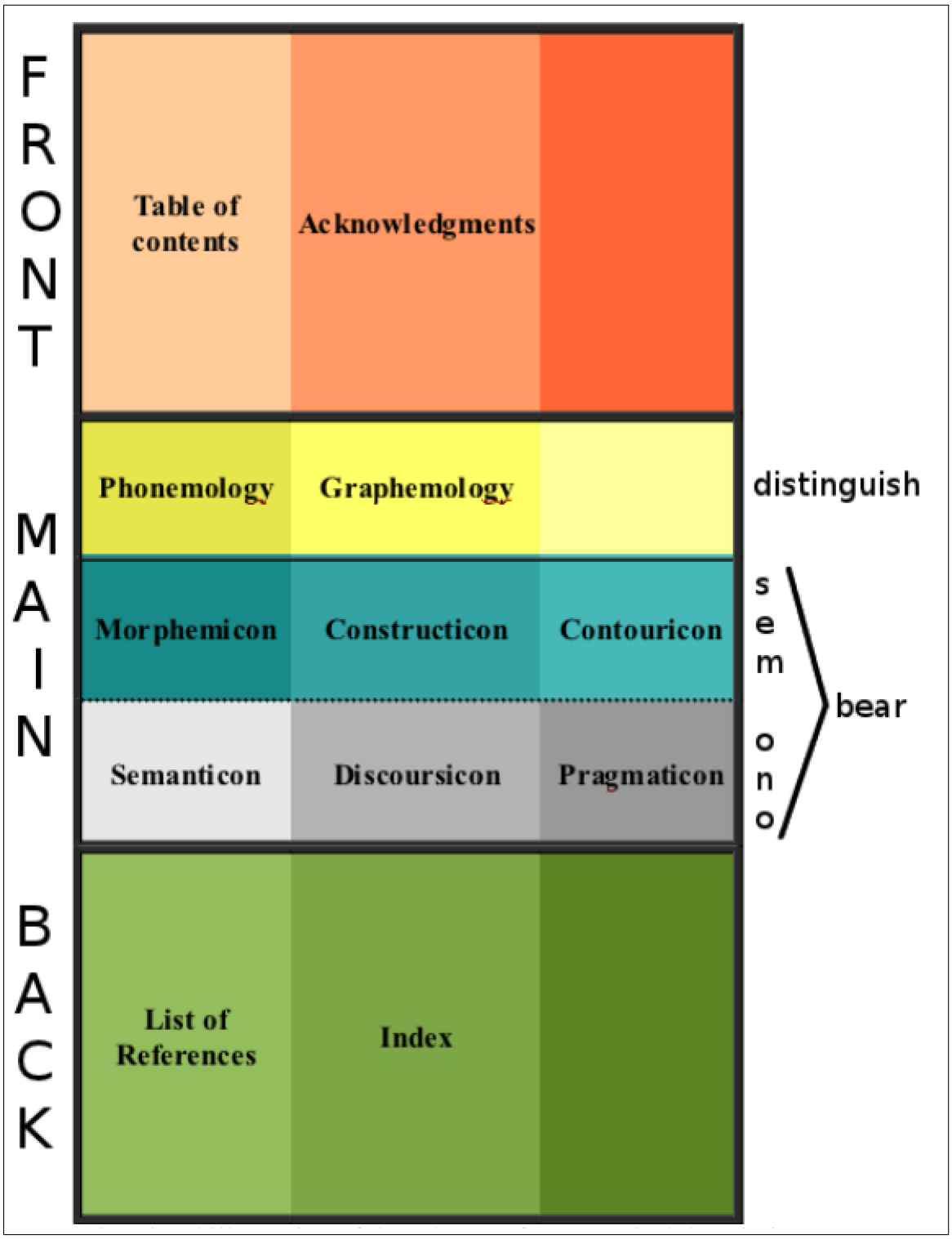


Figure 6.2: Nordhoff's (2012) model of reference grammars

to a section. XLingPaper allows for a terminology, an abbreviation, and a bibliography database in order to manage terms, abbreviations, and references. These databases are not separated from the document, but appear at the end of it. An example database,

however, does not exist, as examples are edited and formatted within the sections.

6.1.5 The Text Encoding Initiative

An XML standard dedicated to the encoding of text is TEI. It was developed to ensure a software- and hardware-independent and sustainable solution for the digital encoding of humanities data. As such, it is widely recognized and used. A great number of academic projects, predominantly situated in the field of digital humanities, are based on TEI.

TEI is an encoding scheme that is defined by the Guidelines For Electronic Text Encoding and Interchange (TEI Consortium, 2020). It encompasses 21 modules that each describe the encoding of specific text genres or textual elements by declaring particular XML elements and attributes. The entire guideline counts almost 600 elements to represent structural units of texts. At the time of publishing, there is no specific module for grammar writing, but existing modules can be used to encode reference grammars. Section 6.5 will propose guidelines for how to use TEI to encode reference grammars. As TEI is constantly expanded to suit the needs of different academic disciplines or text genres, this section will also contain recommendations for new elements and for new structural rules to be added to the format.

For the encoding of a specific document, the user can choose which modules to use and create a schema to validate the produced XML documents. The schema can be specified as a TEI document itself. This schema specification is referred to as the One Document Does it all (ODD). In the ODD document, the user can specify how TEI is customized. For example, each utilized module is indicated. An ODD processor can generate a schema document in one of the following languages: DTD, Regular Language for XML Next Generation (RelaxNG), or the W3C Schema Language. Moreover, such a processor can also create a complete documentation of the customization.

Listing 6.2: The basic structure of a TEI document

```
<TEI>
    <TEIHeader>
        ...
    </TEIHeader>
    <text>
        ...
    </text>
</TEI>
```

Listing 6.2 displays the basic structure of a TEI document. It usually has the root `<TEI>`. TEI not only encodes the text itself but also lets users capture important metadata about the text, the encoding process, and all contributors. This kind of information is stored in the `<TEIHeader>` element, which is the first element within the `<TEI>` root[2].

[2]The composition of the TEI header is discussed in Chapter 2 of the TEI Guidelines (TEI Consortium, 2020)

Below the TEI header, the text itself is located, represented by the element <text>. Which elements are used within <text> depends on the text and its genre. How to encode a grammatical description is discussed in Section 6.5.

6.1.6 TEI-based approaches

TEI has been suggested as the base for a DRG format by Drude (2012, p. 171), as well as Nordhoff and Hammarström (2014, p. 171). Drude (2012, pp. 171-172) merely suggests the use of the TEI framework and the addition of a chapter or sections to the TEI guidelines in order to facilitate the encoding of elements specific to grammatical descriptions. Nordhoff and Hammarström (2014) also suggest using and expanding TEI but discuss specific elements of grammatical descriptions that lack a representation in TEI: named entities, text in the target language, phrase-gloss pairs, linguistic examples, and references.

They use a handful of XML tags for different types of named entities and stress the importance of using identifiers to connect the utilized name to the entity. TEI has, in fact, defined multiple tags for named entities and discusses them in Chapter 13 of the TEI guidelines (TEI Consortium, 2020). Named entities and their encoding in TEI-Grammar are discussed in Section 6.5.3.1.

Nordhoff and Hammarström also use a general tag for text in the target language <objectlanguage> and an associated attribute @iso639-3 that stores the identifier of the language. TEI offers the global attribute @xml:lang to identify the language used in any element. The encoding of text in the target language is discussed in Section 6.5.3.3.

In contrast to the two elements discussed above, phrase-gloss pairs and linguistic examples do lack a representation in TEI. TEI-Grammar proposes different alternatives to encode word-gloss pairs (Section 6.5.3.3) and a whole module for the encoding of IGT and linguistic examples: Text Encoding Initiative Interlinear Glossed Text Format (TEI-IGT) (Section 6.5.5).

Nordhoff and Hammarström (2014, p. 177) mention that references can be separated into references to units within the text and references to units outside the text. While TEI provides the elements <ref> and <ptr> for internal references, external references can be further divided into references to the origin of material, such as a corpus and reference to academic literature. References to academic literature can also be presented by TEI elements, but for the references to the origin of material, they propose the use of attributes with identifiers as values. A detailed discussion about different types of references is provided in Section 6.3.3.4 and the encoding in TEI-Grammar is discussed in Section 6.5.3.4.

Nordhoff and Hammarström's approach also proposes the integration of semantic web relations into the TEI structure. The semantic web and its application in grammaticography is discussed in the following Section.

6.1.7 Good (2012)

Good (2012) proposes another technology as a possible format for DRGs – the semantic web. It draws on the Resource Description Framework (RDF), where different entities are connected through relations between them and ultimately form a huge linked data cloud. The RDF conceptualizes pieces of information as a triplet of what is called a "subject", an "object", and a "predicate". All three parts can be represented through a Uniform Resource Identifier (URI). While the subject and the object are real-world entities such as a specific language, a phoneme, or a lexeme, the predicate represents a relation between the two entities. The combination of multiple relationships between multiple entities can be formalized in a directed labeled graph. Having multiple resources of different grammars, for example, could merge the graphs and create an even bigger connected graph which forms a whole cloud of linked data. The semantic web technologies are used to structure linguistic data (Chiarcos, Hellmann, & Nordhoff, 2011) and are used for a substantial number of projects in the field of linguistics. Good (2012, p. 7) claims that the advantage of the semantic web in contrast to a hypertext approach is the possibility to label relations. This allows for a more profound way of structuring the data. Semantic web technologies will not be used for this project, but could be used to extend the format to link data within grammatical descriptions to external resources. This will be illustrated in Section 6.2.1.

6.1.8 CLDF Format

A more recent approach involves the Cross-Linguistic Data Formats (CLDF) (Forkel et al., 2018), which is a specification to describe different kinds of cross-linguistic datasets, such as wordlists and dictionaries. It is designed for tabular data that is stored in Comma Separated Values (CSV) files. The data can now be described and set into relation with other tabular data using the CLDF ontology that defines relations used for linguistic description. The format intends to build on existing ontologies, such as the GOLD ontology, whenever possible instead of duplicating existing terms. Terms saved in the CLDF ontology are represented as RDF triplets, making use of the semantic web and URIs. This way, the data in each column can be classified and set in relation to other tables.

Florian Matter has developed a workflow to encode a reference grammar with related databases as a CLDF package (Matter, 2019, 2020). The databases include a dictionary and text collection imported from FLEx, an example database, and a bibliography. While these databases are stored in `.csv` files, the grammatical description is encoded as a `.txt` file that is enriched with simple markdown prefixes. For example, a morpheme in the target language can be retrieved from the dictionary using the expression `morph:` followed by the identifier of the morpheme entry. In a similar way, references to bibliography entries, texts, examples, etc. can be marked. This way of encoding the grammatical description and the databases is lightweight and rather simple for linguists to execute. Matter also programmed a couple of Python scripts to display the grammar in a prototype

GET[3].

Matter's approach seems to be a very promising pathway to digital reference grammars. It links data used in the grammatical description to databases such as the dictionary or a text collection. In this way, readers can easily access primary data. The approach includes a prototype GET that should theoretically be able to display different kinds of grammatical descriptions. A GWT is not included but the encoding process is much easier than that of markup languages such as XML. The longevity of the format as well as a possible web application remains uncertain. This problem, however, is also faced by the grammar format and GET proposed in this thesis. This approach and Matter's approach differ in the infrastructure they are embedded into. While Matter's approach utilizes the Cross-Linguistic Linked Data (CLLD) infrastructure, this approach is embedded in the TEI infrastructure.

6.2 Methodology

As this chapter's aim is the development of a format for DRGs, this section needs to discuss the methods used to attain a format that represents the structure of reference grammars. It needs to answer the questions, "Which of the approaches discussed in Section 6.1 will be followed?", "Which technologies are used in the format?", "How has the format been developed?", "How will the format be evaluated?", and "How will the developed format be presented?". These questions will be answered in the following subsections. Section 6.2.1 evaluates the previous approaches and explains which technologies have been used for this approach. Section 6.2.2 discusses the development of the format and how it was evaluated. Section 6.2.3 deals with the presentation of the format.

6.2.1 Technologies

A possible DRG infrastructure (GET and GWT) would require a wide range of different interacting formats (Drude, 2012, p 171). However, there should be one sustainable way to store the grammatical description and to share it with other applications in order to ensure future accessibility of the encoded information. In this way, a standardized method of encoding the grammatical description is indispensable. Only if future DRGs are encoded in a standardized way can they be read and edited by one centrally developed and maintained GET or GWT. Other applications could then rely on the consistency of the data format to be readable into the future. An easy and promising way is to encode the grammatical description in an existing open source format that is widely used and well documented. This way, the format of the grammatical description can build on an existing infrastructure which users will potentially already be familiar with. One such format is XML. XML has often been suggested as the base for a DRG format (Section 6.1).

[3]The prototype GET is accessible at `http://gramr-demo.herokuapp.com/` (retrieved 2021-02-11). The source code is accessible at `https://gitlab.com/florianmatter/gramr/` (retrieved 2021-02-11)

It has the advantages that it is designed to structure textual data in a machine- and human-readable way and that it is widely used. The infrastructure for XML is well advanced and XML technologies such as XML Query (XQuery) and the XSL facilitate the use of XML in applications. Schema languages such as DTD, XML schema, or RelaxNG can be used to define XML based languages – formats which structure specific kinds of data. XML and its infrastructure were discussed in Section 6.1.1

For the encoding of textual data there already exists an XML based language, the Text Encoding Initiative (TEI). As mentioned in Section 6.1, TEI has been suggested as a DRG format by Drude (2012) and Nordhoff and Hammarström (2014). It aims at the semantic encoding of various text genres and should therefore be a suitable language to encode reference grammars in. TEI consists of multiple modules that attempt to cover different text genres. A specific module for grammars does not exist, but guidelines to encode general textual elements, linguistic segment units, and dictionary entries can be found in the existing modules. In this project, TEI has been chosen as the base for the grammar format. The grammar format will consist of guidelines for how to encode reference grammars using TEI and could therefore be regarded as a module of TEI. It will be called, "TEI-Grammar". This format attempts to make use of the existing TEI structure as far as possible but suggests the creation of some new TEI elements to represent units specific to grammatical descriptions. How TEI functions was presented in Section 6.1.5.

The other suggested format approaches (see Section 6.1) can be eliminated for various reasons. PDF and HTML are presentational formats. While they will be necessary formats for the processes of a GET and GWT, they are not meant for extracting and exchanging data. The use of semantic web technologies (Good, 2012; Nordhoff & Hammarström, 2014) and specifically RDF as a format for encoding grammatical information, on the other hand, is a rather complex undertaking. Representing all information stored in a grammar as RDF relations would result in a very formal construct and renders the reference grammar as a pure database. Although this microscopic approach (Section 4.6.4) might be a progressive step towards digitally enhanced language analysis, it could not be a successor of the paper grammar format, as the story would be taken out of the grammar. Moreover, some practical issues seem hard to overcome: Who would encode all those relations? As grammar authors are generally able to edit textual and semi-structured data, they need to be kept away from the encoding of data into the format language. The automatic encoding of textual data into markup languages with the help of templates or layout assistance is a rather easy task. Automatic encoding of textual data into a fine-grained network of semantic relations, on the other hand, seems to be a very complex venture. There are, however, ways to embed some semantic relations into HTML using the technology Resource Description Framework in Attributes (RDFa). This is a method for integrating the advantages of the semantic web into a hypertext grammar. Nordhoff and Hammarström (2014) advocate for the use of unique identifiers within the markup of the grammar to allow for an integration into the semantic web. The problem of encoding the semantic relations, however, is still not solved.

Ultimately, the recommendation or enforcement of a standardized encoding format for reference grammars will probably be up to the institution developing a grammar platform Such a tool and its format could appeal to the majority of grammar authors and become a widely-used standard (Section 5.4.4) (Nordhoff, 2008, p. 319). This dissertation suggests one possible way of encoding reference grammars and illustrates its advantages and disadvantages while developing a GET. TEI-Grammar does not contain a more specialized schema than that of TEI as the exact decision of what should be obligatory, optional, or recommended depends on the operation mode of an editor within the GWT.

6.2.2 Development and evaluation of the format

The development of TEI-Grammar (Section 6.5) involves several steps. First, the structure of reference grammars has been modeled (Section 6.3). Based on the model presented in Good (2004) 6.1.2, the model of this project has been extended with contributions from the literature, findings in other reference grammars, and later ideas that came while encoding parts of the Abesabesi Grammar Sketch (Section 6.4).

The second step involves studying the TEI Guidelines (TEI Consortium, 2020) in order to find existing elements to represent all parts of the model. This resulted in a first draft of the format. The following encoding of the Abesabesi Sketch Grammar aims at detecting missing elements and shortcomings of the first draft. A later draft included adjustments derived from the first grammar encoding. To ensure that the format is applicable to grammatical descriptions of virtually all human languages, the elements of the microstructure and nanostructure have been tested on samples from existing grammatical descriptions. Table 6.1 displays all grammars that have been used to test the format. This grammar corpus has been compiled by selecting random grammars written within this century. They represent languages from different areas and language families of the world. Around a third of these grammars have been accessed as PDFs and the other two thirds as hard copies.

Each element of TEI-Grammar has been tested on a random sample from one or more grammars of this corpus. Because all these grammar samples, together with the Abesabesi Sketch Grammar, have influenced the development of TEI-Grammar it would be difficult to pinpoint which grammar has led to which change in the TEI-Grammar. The grammars are merely listed in Table 6.1 for readers to detect possible biases. Elements of these grammars and the Abesabesi Sketch Grammar will also be used to illustrate elements of grammars that have been modeled (Section 6.3) or encoded (Section 6.5). More comprehensive testing on entire grammatical descriptions would require a longer time frame but could ensure the universal usability of the format. This could be subject to future research.

Finally, an exemplary GET has been developed to test the interaction between the format and an app (Chapter 7). The development of TEI-Grammar has also been a way to test its practical use and to visualize the data encoded in the Abesabesi Sketch Grammar.

Title	Author (Date)	ISO 639-3	Medium
A Grammar of Mani	Childs (2011)	buy	Hard copy
A Grammar of Modern Uyghur	de Jong (2007)	uig	Hard copy
A Grammar of Paluai	Schokkin (2020)	blq	PDF
A Grammar of Sanzhi Dargwa	Forker (2020)	dar	PDF
A Grammar of Semelai	Kruspe (2004)	sza	Hard copy
A Grammar of Southern Min	Chen (2020)	nan	PDF
Bengali	Thompson (2012)	ben	Hard copy
Gramática de la lengua nivacle	Fabre (2016)	cag	Hard copy
Ingush Grammar	Nichols (2011)	inh	Hard copy
Koromu (Kesawai)	Priestley (2020)	xes	PDF
Musqueam Reference Grammar	Suttles (2004)	hur	Hard copy

Table 6.1: Grammar corpus

6.2.3 Presentation of the format

The goal of this chapter is to propose guidelines for the encoding of reference grammars. Along the principles of TEI, which equips the author with a range of structural elements, a schema, and guidelines, this format will contain guidelines for using the existing TEI elements for the encoding of reference grammars without restricting the author in their creativity. The grammar format is meant to provide a method of encoding all basic elements that are common to all reference grammars. On top of this structural base, more elaborate alternatives are suggested, which include a more fine-grained markup. However, these suggestions are more time-consuming and require more consideration regarding how to structure the data. As the format is designed with regards to possible functionalities of a GET (Section 5.4.2), the structural base as well as the alternatives can be read by a GET. However, the more elaborate alternatives allow for the use of additional functionalities and more possibilities for the reader. For example, the linking between examples and audio data is optional but will enable audio embedding under the example in the GET, so that the reader will have the possibility to listen to the actual utterance while studying the example.

The encoding guidelines will be given in Section 6.5. If alternatives to encoding a particular element of reference grammars are available, they will be presented in three levels[4]. While the first level explains the structural base and defines how to encode the element with minimal effort, the second and third levels contain more elaborate methods of encoding the same element. For other elements, only the structural base is presented

[4]These three alternative levels of encoding are not to be confused with the three levels of a grammatical description – macrostructure, microstructure, and nanostructure – presented in the following section.

and optional elements are mentioned. While all levels use existing TEI elements, the higher-level alternatives may include recommendations for a deviating structure of TEI elements and attributes in order to make the encoding of grammars more accurate.

6.3 Modeling reference grammars

This section attempts to deconstruct grammars top-down in order to find the common components that will help us to model the grammar format. The structure will be modeled as a set of elements that are structured hierarchically. This hierarchy can be separated into three levels: the macrostructure, the microstructure, and the nanostructure. Figure 6.3 displays the hierarchical structure of the grammar model and its separation into three levels.

The "macrostructure" comprises the basic unit of the grammar, the grammar entry, and all components above. The "microstructure" counts all components immediately under the grammar entry, such as descriptive prose, lists, or tables. These components will be called, "grammar bricks", as they are necessary to build grammar entries. The composition of these bricks will be referred to as the "nanostructure". It is almost entirely textual data that makes up the nanostructure, but some of the textual elements have special functions, sometimes special formatting, and deserve a separate discussion. These elements will be referred to as "nano-elements". Examples for nano-elements are references, linguistic terms, or words in the target language.

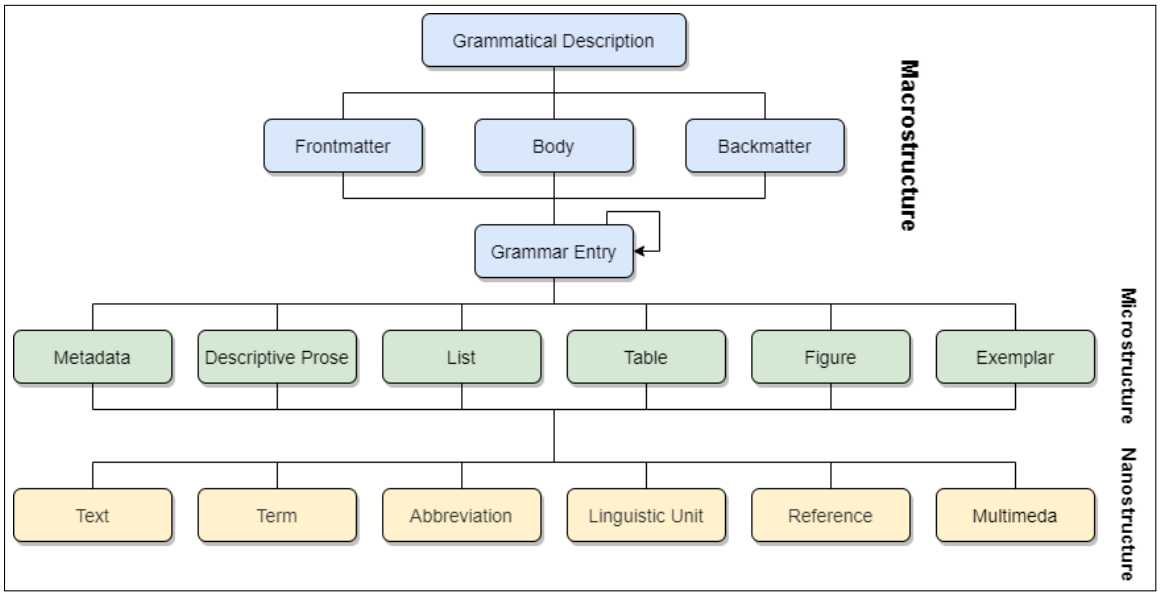


Figure 6.3: A reference grammars formal structure in three layers

6.3.1 Macrostructure

Grammars – in the same way as dictionaries – can be separated into front-, main- and backmatter (Mosel, 2006). Paper grammars usually use this separation in order to isolate

the core descriptions (mainmatter), from navigation tools and introductory chapters in the front- and backmatter, which surround the mainmatter.

6.3.1.1 Frontmatter and Backmatter

The frontmatter and backmatter contain navigation tools such as the table of content (TOC) indices or lists, references, appendixes, and background information about the grammar. Possible parts of the front- and backmatter are contained by the following list. Whether these are parts of the frontmatter or backmatter is decided by the author.

- Frontmatter
  - TOC
  - List of tables
  - List of figures
  - List of abbreviations
  - Acknowledgments
  - Introductory chapters
- Backmatter
  - glossary
  - index
  - list of examples
  - references
  - word lists
  - collection of texts

Introductory chapters The first chapters of a grammar usually give background information about the language, culture, and the grammar itself. Whether these chapters belong to the frontmatter or to the core of a grammar is debatable. While Mosel (2006) counts them to the frontmatter, Nordhoff (2012b, p. 37) treats them as part of the mainmatter. The grammar format should provide flexibility by letting the grammar writer make this decision. Possible content of the introductory chapters is mentioned in Section 4.3.1.

Navigation tools As different readers search for different bits of information in the grammar, it requires tools to help the reader navigate through the text. Some readers might use the grammar as a database and some might prefer reading the grammar cover-to-cover. Both methods require means of navigation. Traditional paper grammars have, at best, two tools for easy navigation: the TOC and the index. While the TOC displays the

organizational structures of the chapters, the index provides an alphabetical list of topics. Both means of text navigation, a TOC and an index, are required in a grammar as they provide different ways to access the text (Noonan, 2005, p. 355). A recurring piece of advice throughout the literature is to not only list the occurring phenomena in this topic list but to also mark the absence of phenomena in the language described (e.g., Noonan, 2005; Cristofaro, 2006). A feature that is expected in a certain area, genetic branch, or that is commonly encountered in natural languages might be absent. If nothing is said about this absent feature, readers cannot assume its absence, even after browsing the whole book. A way to prevent this is marking the absence in the index. Additionally, the absence can be discussed within the mainmatter. The practice of marking the absence of a feature in the index, however, could also be criticized, as the language is checked against a predefined set of features instead of describing it in its own terms. This is the same criticism that has also been held against grammars that strictly follow a maximal outline template like those of the Lingua Descriptive Series (Section 4.3.1).

Additional navigation tools, such as lists of tables and figures, are essentially specialized TOCs. They relate a table or figure identifier (usually numerical) to a page where the particular element can be found within the grammar. The list of abbreviations and the glossary are technically not navigation tools, but tools to clarify terms. Abbreviated forms are expanded in the list of abbreviations and definitions are given to terms in a glossary.

In DRGs, navigation tools are a matter of the GET (Section 7.3.4). The grammar format solely needs to provide ways to encode metadata that can be harvested by navigation tools.

Appendixes Typical appendixes of grammars are a collection of texts and a word list containing vocabulary. These two parts represent the two other pillars of the "Boasian Triade" and were discussed in Section 4.4.

6.3.1.2 Mainmatter

This part of the grammar contains linguistic descriptions and is considered the core of the grammar. The content of the mainmatter is largely influenced by the structure of the language or the language family, current trends in linguistic research, or even by the interest of the author. It was discussed in Section 4.3.1. The mainmatter entirely consists of what Nordhoff and Hammarström (2014, p. 179) call, "structural chapters", which deal with phonology, morphosyntax and other topics that are sometimes referred to as the "core" linguistic topics. Lehmann and Maslova (2004, p. 1872) make further subdivisions with chapters on the "expressive" and chapters on the "significative" subsystems of a language. While the expressive subsystem contains phonology and the writing system, the significative subsystem contains all linguistic signs that have a form and a meaning. A distinction between expressive and significative chapters will not be represented in this grammar format, as the structure of these chapter types does not differ significantly. In

fact, the important difference lays in the linguistic units that are the subject of discussion. These significative linguistic units need to be differentiated from expressive linguistic units.

6.3.1.3 Grammar entries

Grammars usually exhibit a network of chapters and sections that are structured as an inverted tree. The highest level after the root are the chapters. Chapters may contain one or multiple sections, which, in turn, may contain one or multiple subsections. Subsections may also contain one or multiple subsections, and so on. Since "chapters", "sections", or "subsections" are merely labels to indicate the depth in which they are located, I use the term "grammar entry" to subsume all of these labels. Grammar entries can be recursively nested in other entries to assemble the tree-structure. However, this concept does not imply linearity and is therefore also suitable for non-linear grammatical descriptions, which are a network of interconnected grammar entries. The notion of an entry has been inspired by lexicography, from which plenty of parallels can be drawn to grammaticography (Mosel, 2006, p. 50). It is equivalent to the term "annotation" used by Good (2004), who defines a descriptive grammar as "a series of annotations over a lexicon and set of texts".

Whether through nesting or through cross-references, grammar entries need to be interconnected in order to emulate the complexity of a language's structure. In fact, it is rather hard to imagine one entry in isolation without references as hardly any linguistic phenomenon can be described without the support of the description of at least one other phenomenon.. How could negation patterns, for example, be described without referring to their affirmative counterparts? Considering a wider context, the description of phenomena within grammar entries also relies on the discourse about similar phenomena produced within the typological literature or literature on the language family, or linguistic area. A grammar entry can therefore be seen as a grammar unit that is highly interconnected and embedded into not only the grammar itself but into the wider context of linguistic research.

6.3.2 Microstructure

The microstructure of a grammatical description includes all components immediately under the grammar entry. We distinguish the following three formal component types and their manifestations:

- Metadata
  - Identifier
  - Heading
  - Tag

- Descriptive prose
  - Paragraph
- Structured data
  - Table
  - Interlinear examples
  - List
  - Figure

The manifestations of these three types will be referred to as “grammar bricks”, as they are used to construct a grammar entry. Grammar bricks may contain smaller units, which will be discussed as the “nanostructure” of the grammar in Section 6.3.3.

The base of a grammar entry is a coherent text, which alternates with structured data. While the text is meant to give a coherent description of a particular phenomenon, structured data has the task of enhancing the description by visualizing, referencing, or exemplifying data. These two sides of a grammar entry have led to a substantial discussion about the ratio between prose and structured data and, more abstractly, whether a grammar should be treated more like a database (e.g. Cristofaro, 2006) or like a story (e.g. Payne, 2005a, 2014). While Payne (2014) propagates an abundance of text over formalisms and figures as “over-reliance on formulas and diagrams obscure rather than elucidate knowledge” (p. 94), others advocate for plenty of examples and structured description (e.g. Lehmann & Maslova, 2004).

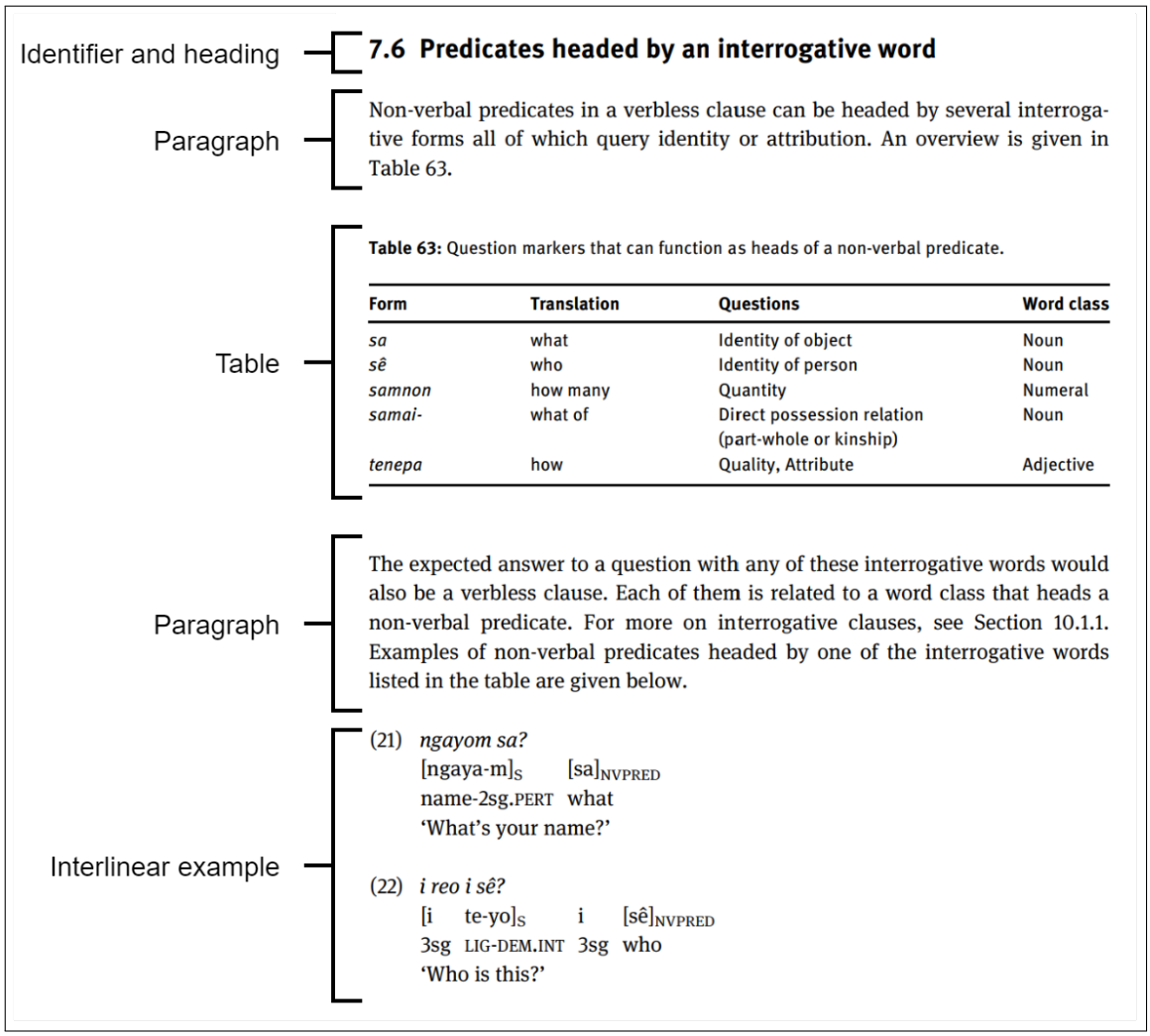
This discussion makes clear that information about language can be broken down into thousands of pieces of structured information, but in order to explain it to an audience, it cannot only be presented in examples and tables, but also needs to be told as a coherent story.

6.3.2.1 Descriptive prose

The object of description in descriptive prose is usually a linguistic feature or a group of features. Like other forms of prose, descriptive prose signals tighter coherence through paragraph units separated from other paragraphs by blank lines or indentation. In grammaticography, these groupings represent the semantic structure of a grammar (Nordhoff, 2012b, p. 41), making paragraphs the central units of descriptive prose. They are situated on the same level as the manifestations of structured data. Figure 6.4 displays a typical grammar entry that contains paragraphs of prose alternating with structured data.[5]

Besides plain text, a paragraph can contain specific elements that have special functions within the grammatical description – the nano-elements (Section 6.3.3. While the plain text has uniform formatting, the nano-elements tend to be formatted differently.

[5]The figure contains an extract of page 231 of the Grammar of Paluai (Schokkin, 2020). The annotations on the left side have been added by the author.



7.6 Predicates headed by an interrogative word

Non-verbal predicates in a verbless clause can be headed by several interrogative forms all of which query identity or attribution. An overview is given in Table 63.

Table 63: Question markers that can function as heads of a non-verbal predicate.

Form	Translation	Questions	Word class
sa	what	Identity of object	Noun
sê	who	Identity of person	Noun
samnon	how many	Quantity	Numeral
samai-	what of	Direct possession relation (part-whole or kinship)	Noun
tenepa	how	Quality, Attribute	Adjective

The expected answer to a question with any of these interrogative words would also be a verbless clause. Each of them is related to a word class that heads a non-verbal predicate. For more on interrogative clauses, see Section 10.1.1. Examples of non-verbal predicates headed by one of the interrogative words listed in the table are given below.

(21) *ngayom sa?*
[ngaya-m]$_{S}$ [sa]$_{NVPRED}$
name-2sg.PERT what
'What's your name?'

(22) *i reo i sê?*
[i te-yo]$_{S}$ i [sê]$_{NVPRED}$
3sg LIG-DEM.INT 3sg who
'Who is this?'

Figure 6.4: Grammar bricks of an entry

Readers are meant to recognize these elements or their function. For instance, bibliographic references have a specific style conventionalized by different research communities. The formatting not only helps readers to recognize the in-text element as a reference but also to identify the particular source document.

6.3.2.2 Structured data

In contrast to descriptive prose, structured data is meant to display data in a purer form. Data is structured to represent relations between pieces of information, to group or classify these relations, or to exemplify parts of the description. There are many ways to structure data. Those that can be pinned down to a specific form will be mentioned in the following paragraphs and later included in the digital data format.

Tables and lists Two very common data structures in all kinds of texts are tables and lists. As they are purely defined by formal criteria independent of their content, they already have standardized representations for digital text formats and can easily be produced in every text editor. Grammaticography uses a few table types that are used for specific content. These table types have only been semi-standardized as they tend to have similar forms throughout different grammars that have been passed down through tradition. Common table types are phonemic inventories or morphological paradigms. Other table types are much more specialized, as they are used for a particular phenomenon that might not be present in every language or they are only used within a specific formal theory.

Interlinear examples The most common data structure in grammatical descriptions, which could be categorized as a table type, is the interlinear example. Before discussing the structure of interlinear examples, an overview about examples in grammars and what forms they have will be given.

Grammatical descriptions rely heavily on examples to illustrate grammatical phenomena, constraints, and rules. Examples come in various forms, only one of them is the interlinear example. Good (2004) distinguishes lexical from textual examples. Lexical examples consist of single words or morphemes that may be accompanied by glosses. Textual examples consist of more complex linguistic units, such as clauses, sentences, or an entire text. Both example types can appear within the prose or as a separate grammar brick. Lexical and textual examples within the prose are nano-elements that will be discussed in Section 6.3.3.1. Figure 6.5 displays lexical examples (*gua*^{3}, *lɯ*^{3}, and *i*^{1}) in a paragraph of the Grammar of Southern Min (Chen, 2020, p. 43). Figure 6.6 displays a textual example in a paragraph of the same grammar (p. 318). The yellow marking has been added by the author to indicate the textual example.

If lexical examples are not situated within the prose, they usually appear in tables or lists. Being separated from the prose, they require an identifier to be referenceable.

The basic pragmatic function of singular personal pronouns in Hui'an is for reference. For example, *gua*^{3} 'I' is used by the speaker to refer to him/herself, while *lu*^{3} 'you' is used to refer to the hearer(s) to whom the speaker is talking, and *i*^{1} refers to the third party including person(s) and thing(s).

Figure 6.5: Lexical examples within the descriptive prose

Example (10) implies that both the comparee (*gua*^{3} 'I') and the standard of comparison (*i*^{1} 'she') are tall, based on which, the comparee is taller than the standard of comparison. However, without *ko?*^{7}, i.e. *gua*^{3} *pi*$^{3-2}$ *i*^{1} *kha?*$^{7-8}$ *kuin*^{2} 我比伊恰懸 'I am taller than her' does not imply that both the comparee and the standard of comparison are tall. In other words, it could be that both the comparee and the standard of comparison are short, while the comparee is nevertheless still taller than the standard of comparison. Note also that the English translation provided above for (10) is not semantically equivalent to the comparative construction of the Hui'an dialect, since the English sentence 'I am even taller than her' does not necessarily imply that both the comparee and the standard are tall.

Figure 6.6: Textual example within the descriptive prose

Figure 6.7 displays lexical examples in a table within an extract of "A grammar of Sanzhi Dargwa" (Forker, 2020, p. 107).

Because readers that do not understand the target language might find it hard to follow more complex examples, these tend to not be situated within the prose but as a separate grammar brick, the interlinear example. Interlinear examples not only provide a translation but also glosses for each word or morpheme.

All example types can be used alone or in a set with other examples. Prose examples in sets are usually simply listed and separated by commas. Free lexical examples are rarely found alone, as the reason for separating them from the prose is often that multiple examples are hard to integrate into the text (like in Figure 6.7). Sets of lexical or textual examples can have an identifier as a whole or sub-identifiers in order to reference any of

Table 4.10: Interrogative pronouns

ča	'who'	*ceʁuna*	'which'	*ceq:el*	'when'
ce	'what'	*kut:i*	'which'	*čujna*	'how many times'
čina	'where'	*cel*	'why'	*kusa*	'how much'
cet'le	'how'	*celij*	'why'	*čum*	'how many'

Figure 6.7: Lexical examples in a table

the examples in particular.

An overview of the four example types in grammatical descriptions and their formal manifestations can be seen in Table 6.2.

	Within prose	Independent grammar brick
Lexical	Linguistic unit	List or table
Textual	Linguistic unit	Interlinear example

Table 6.2: Different example types and their formal manifestations

Having differentiated the various forms of examples in grammatical descriptions, we can now focus on the particular form this section is about – the interlinear example. As mentioned before, the interlinear example is a structure in which textual examples are presented outside the prose that facilitate understanding of the structure of an utterance. Before analyzing this form, we need to make some important distinctions and clarify some terms that will be important later in the discussion.

In most cases, textual examples stem from recorded utterances, unless they stem from non-recorded elicitation sessions. Moreover, they can be produced through introspection, which is a rather controversial method in descriptive linguistics. In a language documentation setting, recorded utterances are transcribed (for example in an ELAN[6] file or as a spread sheet). Transcribed utterances might also be translated and glossed at this stage. The sum of all transcribed utterances that are chosen to be used for a grammatical description will be called the "text collection", whether the collection is actually published or not. A text collection might also include written primary data. In this case, it could be better to call the segments "sentences", but for simplicity reasons, I will also include them in the term "transcribed utterances". In order to become an example for a specific phenomenon, the transcribed utterances are modified and formatted. Modification includes shortening the utterance and deleting optional components so that they do not distract the reader. Formatting includes highlighting components by using bold script, indicating relations with brackets, arrows, or indexes, and marking zero elements by leaving a gap or using the empty set element (∅). Good (2004) makes a further distinction between "examples" and "exemplars". Examples can contain multiple phenomena and therefore be an example for all of these phenomena. Good refers to an example that was chosen to exemplify a specific phenomenon as an "exemplar". This distinction is rather abstract, as an exemplar is not a formal category but rather a relation. An example is thus an exemplar for a specific phenomenon. This also means, that every example being used in a grammatical description is an exemplar for the phenomenon that is described in the same grammar entry. It might, however, also be an example for other phenomena, even though it is not modified and formatted to specifically illustrate those phenomena. This principle has been used by Haspelmath (1993, pp. 530-538) in his Grammar of Lezgian, who provides cross-references between exemplars and additional examples. He lists all

[6]retrieved 2021-03-18, from`https://archive.mpi.nl/tla/elan`

exemplars and links them to other examples in the grammar that contain the same phenomenon. The distinction between exemplars and examples will play a further role in the suggested data format, where examples are stored in an example-database and only their instantiation within the grammatical description will be referred to as the exemplar.

Summing up, the following stages from a recorded utterance to an exemplar have been distinguished:

RECORDED UTTERANCE

→ TRANSCRIBED UTTERANCE

→ EXAMPLE (MODIFIED AND FORMATED UTTERANCE)

→ EXEMPLAR (EMBEDDED EXAMPLE)

Interlinear examples are examples that come in a text format, which is generally referred to as IGT. While standardization has been attempted by different scholars (e.g., Lehmann, 1982; Comrie et al., 1982; Drude, 2003), the IGT format is generally based on certain glossing conventions. Editing and printing do not pose many formal limits for the layout of examples, and authors have therefore used their creativity to come up with different methods formatting an example to best illustrate their point.

The base form, which has been propagated by (Lehmann, 1982) and (Comrie et al., 1982), consists of three tiers, such as the formalized example by Weber (2005a, p. 451):

(Number) *Morphemes*
GLOSSES
'Translation'

This layout includes the modified utterance separated into morphemes in the first tier, the glosses aligned with the respective words in the second tier and a free translation in the third tier. According to the widely accepted Leipzig Glossing Rules (Comrie et al., 1982), words in the first two tiers are separated by blank spaces, while morphemes within a word are separated by hyphens or the equals sign (=) in case of a clitic. A numeral identifier is usually placed to the left of the example to make it referenceable. This base structure can be seen in Figure 6.8, which displays an example from the Koromu Grammar (Priestley, 2020, p. 71).

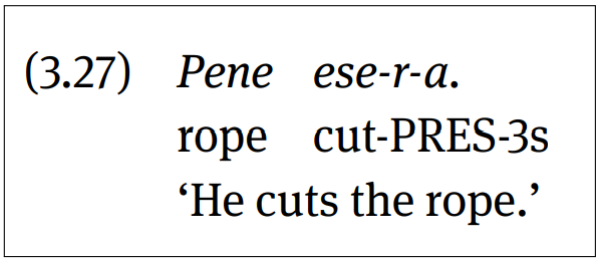

(3.27) *Pene* *ese-r-a.*
rope cut-PRES-3s
'He cuts the rope.'

Figure 6.8: Lexical examples in a table

Depending on the needs of the author and the nature of the language, the three-tier structure may be extended to include, for instance, orthographic writing (Weber, 2005a, p. 452), suprasegmental processes (McGill, 2007), or to include multiple layers of description (Drude, 2003).

Authors use quite an amount of more or less conventionalized means to highlight different aspects, push information into the background, or annotate relations or categories (Weber, 2005a). Frequently used means to express these functions include formatting, such as italics, boldface, underlined letters, or capitalization, brackets to indicate complex syntactic units, and indexes to mark coreference. It will be essential to offer a wide range of these means to the author of a DRG.

Besides its own (numeral) identifier, interlinear examples often include a way to identify where the utterance can be found in a recording. This identification provides readers with the possibility to compare the recorded utterance with the given transcription and to engage more directly with the author's claims. It also allows for an investigation of the context of the utterance and the metadata of the recording. Both are important, as grammar can be dependent on an utterance's environment, the genre, or the speaker. Tracing the original utterance, the utterance's context, and the metadata back through the utterance's identifier, however, is a rather time-consuming process. A DRG could provide faster access through hyperlinks.

Figures Apart from all those data structures that can be subsumed under tables and lists, there are many more data structures that come in very different forms and would be hard to standardize. These include all kinds of diagrams and figures, for example, syntax trees, autosegmental annotations, or other theory-driven formalisms. A standardization of all these different formalizations is not a reasonable undertaking, which is why all of them will be subsumed under the term "figure" within this section. Moreover, complex formalisms, especially when they are part of a contemporary linguistic theory, are often advised against. The argument is that they distract the reader from the information presented rather than helping them understand it. Problems that arise with the use of theory include the potential to become outdated (Payne, 2014, p. 94) and the inaccessibility of the descriptive information for people who are not familiar with the theory in question (Rice, 2005, p. 397). However, linguistic theories can help explain certain phenomena and make a grammar more coherent (Rice, 2005, p. 403). Rice (2005) suggests that grammar writing should be informed by theory, but it should not be a topic of the grammar.

6.3.2.3 Metadata

An entry's metadata consists of the heading, an identifier, and tags. In paper grammars, headings are printed at the top of a grammar entry, usually preceded by the numeral identifier of the grammar entry. This identifier often consists of numbers separated by dots to mark the location of the entry within the grammar's tree structure. Entry number

“4.2.3”, for instance, represents the third subsection of the second section of the fourth chapter. The third type of metadata found in paper grammars are tags. Although often not visible in the grammar entry, they are existent as indexes. An index lists keywords that are related to certain pages or sections of the grammar. In turn, this means that grammar entries or pages can be tagged with one of the index keywords. As pages will not play a role in digital grammars, tags will only be attributed to grammar entries.

6.3.3 Nanostructure

The nanostructure includes all components that the grammar bricks of the microstructure are made of. Next to plain text, which will not be discussed in detail, the nanostructure consists of other textual and non-textual components that have special functions and require specific formatting: The nano-elements.

6.3.3.1 Linguistic units

Unless it is a monolingual grammar, a grammar contains text in at least two languages: The metalanguage and the target language. While the metalanguage is the language that the description is written in, the target language is the subject of the description. Units of the target language are used throughout the description as labels for the phenomenon described or for exemplification. It is important to differentiate text in the metalanguage from text in the target language. Otherwise, readers might try to parse a unit of the target language as part of the sentence it is embedded in, in the metalanguage. This is why paper grammars usually differentiate the target language by formatting it differently. A common way is to use italics. Target languages with another script, on the other hand, do not require different formatting, as they are already visibly distinguishable from the metalanguage.

There are linguistic units for several linguistic levels that can occur in grammatical descriptions. The following list contains units that occur in existing grammatical descriptions. A detailed distinction of these categories will be undertaken in the format for DRGs (Section 6.5.3.3).

- expressive units
  - (allo-)phone
  - phoneme
  - grapheme
  - syllable
- significative units (form side)
  - (allo-)morph
  - morpheme

  - lexeme
  - word form
  - construction
  - intonation contour
  - phrase
  - clause
  - sentence

- significative units (function side)
  - meaning
  - grammatical categories

Linguistic units can be grouped into expressive and significative units (Lehmann & Maslova, 2004, p. 1872). Expressive units are those that distinguish meaning, and significative units are those that bear a meaning. Paper grammars usually distinguish different expressive units via formatting. Established formatting devices are slashes for phonemes (e.g., /f/), square brackets for phones (e.g., [f]), and angle brackets for graphemes (e.g., <f>). The use of these distinguishing devices is important for a reader's comprehension of which linguistic level the current description is referring to.

As significative units have a form and a function side, authors have multiple ways to refer to them within the grammatical description. They might only use the form side as a label (6.1 a), a combination of form and function (6.1 b), or only the function (6.1 c).

(6.1)
(a) In this case, the particle *gí* is situated between the focus fronted phrase and the rest of the clause.

(b) In this case, the particle *gí* 'FOC' is situated between the focus fronted phrase and the rest of the clause.

(c) In this case, the focus particle is situated between the focus fronted phrase and the rest of the clause.

Example 6.1 displays three versions of a possible sentence of a description for Abesabesi illustrating the three ways to refer to a significative unit. While the combination of both sides is the most accurate way to refer to a linguistic unit, the other options are often used after the significative unit has already been introduced. However, most readers are not familiar with the target language and only using one side of a significative unit might not be sufficient most of the time.

A common way to refer to the function side of a significative unit is to present it in quotation marks. Grammatical categories are usually represented as abbreviations in small capitals (Comrie et al., 1982). A distinction between different significative units through formatting, however, is not common. It can only be derived by context. Affixes,

for example, are indicated using a hyphen (e.g., -ed for the English simple past suffix). However, not all morphemes are affixes. A distinction between lexemes and word forms is usually completely lacking. Clear distinctions of linguistic units and their benefits for DRGs will be discussed in Section 6.5.3.3.

6.3.3.2 Terms

Descriptive prose and other grammar bricks are interspersed with terminology that pinpoints specific topics of linguistic theory. Conventionalized terms ensure that readers understand a description as it was intended by the author. Yet, terms may not refer to exactly the same phenomenon across different theories. Good (2004) therefore suggests that terms used should always be connected to a specific ontology.

But what exactly are ontologies? Within this thesis, they are regarded as a structured arrangement of terms essential to a specific field of study. Ontologies are important, as the definition of a term in isolation usually lacks clarity. In order to decode a term used by a grammar writer, we need to have it not only defined, but also differentiated from or related to other terms. Ideally, ontologies comprise all terms that are necessary to describe data within a field of study, they provide definitions that distinguish a term from others and they set all terms into a structure that sets the concepts into relation with each other.

Good (2004) distinguishes three types of ontologies from which grammar writers draw their terminology: “general ontologies”, “subcommunity ontologies”, and “local ontologies”. General ontologies contain terms that are assumed to be used or at least understood by all linguists. An example is BLT, which is the base for terms adopted by many typologists and descriptive linguists. General linguistic ontologies also exist digitally – some approaches being GOLD (Farrar & Langendoen, 2003), isoCAT (Snijders, Windhouwer, Wittenburg, & Wright, 2009), OLiA (Chiarcos & Sukhareva, 2015), and OnLiT (Lehmann et al., 2017). Although general ontologies are supposed to cover terms generally used in linguistics, they might be insufficient when it comes to phenomena unique to a language family that have already been described in a long areal tradition. Subcommunity ontologies, while drawing on terms of general ontologies, are understood by a subcommunity of linguists. These subcommunities might be linguists interested in a certain language family or a certain grammatical framework.

All other terms that do not fit into these two types of ontologies belong to a local ontology. These terms might be used to subdivide categories drawn from a general ontology (Good, 2004, Section 1.1) or they may be used simply because a language exhibits a category that has never been described before (Pawley, 2014, p. 97). However, the latter case requires a lot of care. One might have just not heard about a category, although it has indeed been labeled before. Given the diversity of grammatical structure throughout the world’s languages, Rice (2005, p. 308) points out that grammar writers tend to see unique traits in a language’s features rather than focusing on the regularities. This urges them to use an ideosyncratic label rather than a common one that is justified

by the feature's regularities. Many scholars stress the importance of thorough preparation by obtaining a good knowledge of typology (Noonan, 2005; Cristofaro, 2006). The scale between universalism and particularism introduced in Section 4.2 brings with it the problem of choosing between an established term and a new, unique term. Even if terminology from general ontologies is used, the label given to a language-specific category is likely to not match the definition of the term completely. This is why the used terminology must always be explicitly defined (Noonan, 2005, p. 358). Cristofaro (2006, p. 156) additionally recommends mentioning the criteria used to identify the category in order to make a term comparable. Using general terms for language-specific categories entails the risk that language-specific categories cannot be distinguished from "universal" categories. Haspelmath (1993), following a practice used by Comrie (1976), distinguishes categories specific to the Lezgian language by capitalizing the labels.

Summing up the recommendations results in a work flow for the use of terminology: 1) obtain a good general knowledge about typology and the kind of phenomena you want to describe 2) try to use terminology from a general ontology 3) whenever this will not help you, use terminology of a subcommunity ontology 4) if the feature is not comparable to any other phenomenon described in the literature, use your own terminology 5) define your terminology, no matter where you took it from. The recommendations collected above, however, stem from typologists, who work with grammars to compare phenomena and who rely on consistent terminology. As the work flow is based on these recommendations, it is merely a guideline for achieving consistent typology beyond a subcommunity. In practice, many grammar authors prefer to use subcommunity terminology over general terminology or tend to coin their own labels. For DRGs that include a grammar writing tool, this leaves two options. Either the authors will be guided to use consistent comparative terminology, or they will be left free to decide which kind of terminology they use, as long as they reference theoretical resources where the terms are defined or define them themselves.

Taking a more formal perspective, it can be said that terms each comprise a label and a gloss and have a fixed hierarchical position within an ontology. The label consists of one or a few words that are used within the grammatical description to reference the term within the ontology. A gloss contains a definition of the used term. Glosses are not given within the prose. In paper grammars, they can be collected in a glossary, where each term is listed with its respective gloss. The grammar format thus needs to emulate the structure of an ontology and provide a term element within the grammatical description that references terms from the ontology.

Besides technical terms, as discussed before, terms can also reference concrete entities in the world. These "named entities" are not distinguished in paper grammars. In computational linguistics, however, resolving named entity references is an important field of research, as it allows for the identification of entities and the creation of data networks. The encoding of named entities will play a bigger role in the grammar format (Section 6.5.3.1).

6.3.3.3 Abbreviations

Like other forms of prose, grammatical descriptions make use of abbreviations. Terms can be abbreviated and, in particular, long terms can be written as acronyms or other types of abbreviations to save time writing and, ideally, to hasten the reading flow. Abbreviations refer to expanded terms, which are usually mentioned the first time an abbreviation is used, like in the following example: vowel harmony (VH). From now on, abbreviations are used without the expanded term. List of abbreviations exist in the front- or backmatter to give the reader the possibility to look up abbreviations and the term they stand for.

A particular type of abbreviation in linguistic descriptions is the grammatical gloss. Interlinear examples (Section 6.3.2.2) contain a tier of glosses where language-specific grammatical categories are abbreviated and written in small capitals. Example 3.68 from the Abesabesi Sketch Grammar has been repeated here as Example 6.2. In its gloss line, the abbreviation PRF is used to refer to the language-specific grammatical category "perfect".

(6.2) *ɔban èdɛ́n ka ga*
child that PRF be_big
'the child has become big/is already big' (ibe117-00.134)

Besides grammatical glosses, grammatical descriptions also contain abbreviations of technical terms and named entities. The set of abbreviations is thus often a subset of the terms used in a grammatical description.

6.3.3.4 References

Good (2004, Section 1.3) distinguishes four types of references typically found in grammars: references to lexical items, to exemplar data, to terms drawn from ontologies, and to other sections. For the purposes of this project, a few comments and adjustments will be made to this list.

Firstly, a grammatical descriptions usually includes various items of the target language (Section 6.3.3.1). Lexical examples that are used within the prose, such as morphemes or words, can be understood as a reference to their respective lexical entries, as they are usually given in their lemma form. If they appear with their gloss, they might even refer to a specific sub-entry. This abstract reference could in DRGs be extended to a concrete reference by means of hyperlinks. Besides formatting a linguistic unit to indicate that it contains text in the target language, the element could be clickable and direct the reader to the respective lexical entry in a dictionary. This way, additional data of a lexical entry could be looked up in a very short time.

More complex linguistic units, such as clauses or sentences, are usually taken from actual recordings. Interlinear examples, in particular, tend to contain a reference to the corpus data in the form of an identifier. Example 6.2, for instance, contains the identifier of a recording segment within the Abesabesi Corpus "ibe117-00.134", to the right of the free translation. Using this reference, readers can look up the file in the corpus and

retrieve the context and metadata. A DRG could also create hyperlinks to the corpus files in order to guarantee fast access to the original utterance with its context and the metadata.

Secondly, references to exemplar data and to other sections are both references to other elements within the grammar. We could likewise include references to other structured data such as tables, lists and figures.

Thirdly, abbreviations also refer to terms drawn from ontologies. Instead of referring to a term and its explanation, they use an abbreviated form of a term and refer to the full form of the term.

Lastly, another type should be added, which is essential to every academic study: references to other publications. In grammaticography these might be references to previous descriptions of a phenomenon, to typologically similar data of other language descriptions, to websites containing databases, or to works about linguistic theory that support the description.

Considering these four points, we result in a complete list of five reference types:

- References to lexical entries
- References to corpus data
- References to other elements within the grammar
- References to terms drawn from ontologies
- References to other publications

References follow specific patterns in grammatical descriptions to be identified as such. Lexical entries are usually referenced by their orthographic representation in italics followed by a gloss in apostrophes: *orthographic_representation* 'gloss' (Good, 2004, Section 1.3). The orthographic representation can be replaced by a phonetic or phonemic representation. When using different representations within the grammar, it is important to signal which kind of representation has been chosen. This can be done by different kinds of brackets or different fonts. As mentioned in Section 6.3.3.1, linguistic units in the target language need to be formatted differently from the descriptive prose. Otherwise, it is hard to process a switch from the metalanguage to the target language.

Corpus data can be referenced using the identifier system of the corpus, as shown earlier in this section. The identifier system of the corpus is usually defined by the researcher that compiled the corpus. It needs to direct to a specific file and ideally to a specific segment of the file.

Structured data, such as interlinear examples, can be implicitly referenced, if they follow the descriptive prose directly. More often, however, the grammar brick is given a unique identifier, i.e., a number that can be used to reference it within the prose. Other grammar entries are referenced by unique identifiers that are typically numeric values separated by dots to symbolize the hierarchic structure of the grammar entries.

As numeric values of other grammar entries or other grammar bricks are usually only unique within their category, a clear reference needs to also include the type of the component, for instance “Section 2.4”, “Example 2.4”, or “Figure 2.4”.

Terminology is often used without referencing an ontology from which it is drawn. Some grammars distinguish terminology by applying typographical conventions such as italics or capitalized letters. The formatting of abbreviations has already been mentioned above in this section.

The last kind of reference, a reference to other documents, has the same format as in other scientific prose: Depending on the citation style, the original document is referenced by a combination of the author and date or gets another kind of unique identifier. More metadata about the source is given in the references section of the backmatter.

6.3.3.5 Multimedia

Audio and video are multimedia types that cannot be embedded in paper grammars, which are bound to the book format. DRGs, however, could make use of these multimedia types in order to support the analysis and make the research verifiable (Section 5.2.1). Pictures, on the other hand, are a multimedia type that is supported by the book format and have been used extensively. A picture can be part of the grammar brick “figure”. A figure is not the same as a picture, as a figure contains one or more pictures as well as other possible elements, such as a caption and an identifier.

6.4 Encoding the Abesabesi Sketch Grammar

The processes of developing TEI-Grammar and encoding the Abesabesi Sketch Grammar took place simultaneously with one process nurturing the other. While the encoding process relied on the encoding standard, the development of said standard was constantly tested and refined using the experiences had while encoding the sketch grammar. Additionally, a random set of sample elements from the grammar corpus have been encoded (Section 6.2.2). However, it is important to note that the experiences had while encoding the Abesabesi Sketch Grammar have been the biggest source of inspiration for TEI-Grammar. Although the structure of other grammars has been considered, TEI-Grammar might still be biased by the structure of Abesabesi and my personal style of writing grammars.

The encoding of the Abesabesi Sketch Grammar was undertaken “manually”, as no GWT has been developed to facilitate the encoding. This means that only an XML[7] editor was used to type all grammar documents. Therefore, the encoding process has been rather time-consuming. Instead of encoding the entire sketch grammar, twelve grammar entries have been selected from different parts of the the grammar to ensure that as many different structural elements of a reference grammar as possible are included.

[7] I have used the Oxygen XML Editor (retrieved 2021-02-02, from `https://www.oxygenxml.com/`)

Additionally, the entire structure of the sketch grammar has been encoded by creating empty grammar entries with headings in a linear manner. This has been done in order to display the entire structure in the TOC.

The XML editor also facilitated the encoding by validating the document against a TEI schema in order to ensure that the grammar document was well-formed. Nevertheless, the method of manually encoding large quantities of text and other data is still error-prone. This is because the presented guidelines have not been fixed in a schema, as they are only meant to be recommendations. Yet, the encoding process primarily served to detect shortcomings of the format in order to refine it. The development of a GET that can read the format and display the grammar document as hypertext pages was then a second step to detect shortcomings of the format (Section 7).

The encoding of the Abesabesi Sketch Grammar involved the main grammar document and all seven databases. Three databases were produced by transforming existing data into TEI. The text database and the lexical database were transformed from FLEx output and the bibliography database was transformed from Mendeley output. The terminology, abbreviation, and example database were encoded manually.

Only valid TEI encoding was used. Section 6.5.6 will propose alternatives that are not TEI compliant. These alternative elements have not been used for the encoding of the Abesabesi Sketch Grammar. Encoding examples of the Abesabesi Sketch Grammar will be given throughout the following section where TEI-Grammar is presented.

6.5 Encoding reference grammars

While a model for reference grammars has been formed in the previous section, this section aims at transferring the model into a TEI serialization and presenting a method to digitally encode reference grammars. It will be structured like the previous section and deal with all three layers of a grammatical description: the macrostructure, the microstructure, and the nanostructure. These layers will all be encoded on what has earlier been referred to as the “main grammar document”. This document is a digital file with the base structure of a TEI document (Section 6.1.5).

Section 5.4.1 proposed a separation between the grammatical description itself and the structured data. Following the approach, TEI-Grammar can also be used to encode the seven proposed databases in separate TEI documents. Not all of these databases need to be encoded to have a functioning grammar. For instance, a dictionary database is not required to write a grammatical description, as words or other units in the target language do not have to be connected to a database. However, a resulting grammar will lack some of the presented functionalities if it lacks a database. In this case, words or other units in the target language cannot be clicked to guide the reader to the lexical entry in the lexical database. However, this section will sketch out the maximal process of encoding the grammar while indicating alternatives for leaving out certain parts of the encoding.

This maximal process of encoding comprises a total of seven digital documents[8]. Other resources that are part of the grammar are multimedia files that can be referenced inside the seven text documents.

1. The main grammar document containing the grammatical description and referencing entries of the databases
2. The terminology database containing terms and their definitions, hierarchically structured as one or multiple ontologies
3. The abbreviation database containing abbreviations and their expansions
4. The bibliography database containing bibliographic entries referenced in the main document
5. The text database containing texts (ideally glossed and translated)
6. The example database containing example utterances taken from the texts and edited for didactic purposes
7. The lexical database containing lexical entries

While the following three subsections deal with the three layers of the main grammar document, later subsections discuss the encoding of the database documents. As mentioned in Section 6.2.1, TEI-Grammar aims at the encoding of grammatical descriptions using existing elements. A final subsection (Section 6.5.6), however, will also propose alternative elements that do not exist in TEI yet that could better represent units specific to grammars

6.5.1 Macrostructure

A separation into front-, main-, and backmatter is provided by TEI's `<front>`, `<body>`, and `<back>` elements. Navigation devices, such as the TOC, the index, and lists of tables/figures/etc. can be computer-generated. The placeholder element `<divGen>` can be used to mark the space where the particular navigation device should be generated when rendered linearly (for example, in a print copy). The `@type` attribute is used to indicate which navigation tool or list should be generated at the location of the placeholder. If the GET offers linear outputs of the grammatical description, it needs to have a controlled vocabulary of `@type` attribute values, such as "`tableOfContent`", "`index`", or "`listOfAbbreviations`". This is necessary for the application to understand what exactly should be generated. TEI does not impose any restrictions on the `@type` values. A combined grammar platform, as described in Section 5.4.4, could ensure the coherent use of the values. The author would only have to decide where the particular navigation tools and lists would be situated and the GWT would produce the `<divGen>` elements and

[8]The list is repeated from Section 5.4.1

use the controlled vocabulary that can be read by the GET. The `<divGen>` element can be used within the frontmatter, the backmatter, and theoretically also in the mainmatter (Listing 6.3).

Listing 6.3: Frontmatter

```
<front>
        <divGen type="tableOfContent"/>
        <divGen type="listOfFigures"/>
        <divGen type="listOfTables"/>
        <divGen type="listOfAbbreviations"/>
</front>
<body>
        ...
</body>
<back>
        <divGen type="index"/>
        <divGen type="listOfExamples"/>
        <divGen type="listOfTexts"/>
        <divGen type="listOfWords"/>
        <divGen type="references"/>
</back>
```

The tree structure of grammar entries can be represented through nested `<div>` elements within the body. If a distinction between chapters and sections is preferred, this can be declared in the `@type` attribute: `<div type="chapter">`. This, however, is not necessary as the distinction can be generated by an application. First-level `<div>` elements are chapters and lower-level `<div>` elements are sections. `<div>` elements should have an identifier in order to allow for referencing. The identifier needs to be unique and is saved in the `@xml:id` attribute. Additionally, the `@n` can be used to store the section's position within the text. These position numbers are typically printed before section headings in paper grammars and help the readers to identify the position of a section within a guided path through the grammar. Position numbers do not suffice as identifiers themselves because other structural elements, such as tables or examples, might also be numbered. An identifier therefore needs to be the combination of element-type and position number (e.g., "ch1" for the first chapter or "tab15" for table number 15).

Listing 6.4: Chapter structure

```
<div xml:id="ch7" n="7">
        <head>Word Formation</head>
        <div xml:id="ch7-sc1" n="7.1">
                <head>Deverbal Nominalization</head>
        </div>
```

```
        <div xml:id="ch7-sc2" n="7.2">
            <head>Composition</head>
        </div>
        <div xml:id="ch7-sc3" n="7.3">
            <head>Reduplication</head>
        </div>
</div>
```

6.5.2 Microstructure

The microstructure is the structure within a grammar entry. Each entry needs to have a title and an identifier and can contain tags or indices. These elements make up the metadata of an entry, which feeds the navigation tools and facilitates the search for a specific topic. The rest of the microstructure consists of the grammar bricks discussed in Section 6.3.2: descriptive prose and structured data.

6.5.2.1 Metadata

Each grammar entry must have a title and an identifier and may have metadata tags, such as keywords, that will feed the index or values for quality assessment and relative importance. While the identifier is represented as an attribute of the `<div>` element, a grammar entry's title is represented by the `<head>` element, which can be situated within the `<div>` element. Tags for other metadata, such as quality assessment and relative importance, have been explained in Section 5.2.1 (Maxim 13). TEI provides the `<note>` element, which can be used as a tag. TEI-Grammar uses the attribute `@ana` to indicate which property is tagged and the content of the element to indicate the value. A value could be a numeric value or text data. The data type of the value determines how it will be rendered in an application. For instance, if numeric values between 0-100 are used, the GET could read the value as a percentage to which a bar is filled with a specific color (see Figure 6.9).

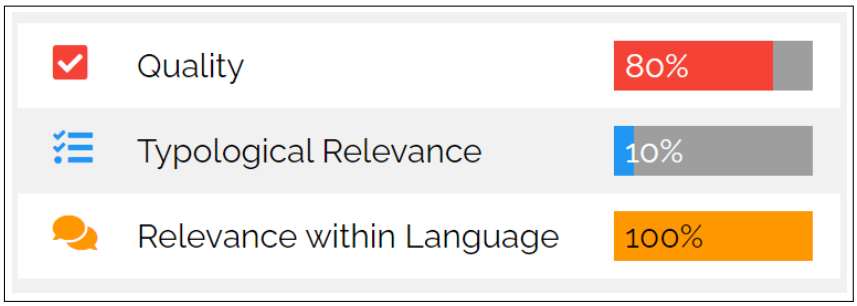


Figure 6.9: Tags for quality assessment and relative importance

For tag keys (indicated in the texttt@ana attribute) the Abesabesi Sketch Grammar used `#qual` for quality, `#typRel` for typological relevance, and `#langRel` for relevance

within language. These abbreviated values should be decoded in an <interpGrp> element at the end of the <body> element, as shown in Listing 6.5.

Listing 6.5: Tags and their interpretation

```
<body>
        ...
        <div xml:id="ch2-sc4-4" type="section" n="2.4.4">
                <head>Vowel harmony</head>
                <note ana="#qual">80</note>
                <note ana="#typRel">10</note>
                <note ana="#langRel">100</note>
        </div>
        ...
        <interpGrp>
                <interp xml:id="qual">Quality</interp>
                <interp xml:id="typRel">Typological relevance</interp>
                <interp xml:id="langRel">Relevance within language</interp>
        </interpGrp>
</body>
```

This method of tagging a grammar entry could also be used for indicating the domain and priority of a grammar entry in the dynamic path approach proposed by Nordhoff (2008, p. 315) (Section 5.3.7).

Each grammar entry can also have certain keywords attributed to it in order to facilitate keyword-based searches and to feed the index. Keywords are stored in <term> elements, which are grouped together in the <index> element. As the used keywords are drawn from an ontology stored in an external document (see Section 6.5.4.3), the <term> elements are empty but refer to the respective term within the ontology through the @ref attribute. Listing 6.6 shows a grammar entry with the whole range of metadata.

Listing 6.6: The index

```
<div xml:id="ch2-sc4-4" type="section" n="2.4.4">
        <head>Vowel harmony</head>

        <note ana="#qual">80</note>
        <note ana="#typRel">10</note>
        <note ana="#langRel">100</note>

        <index>
                <term ref="terminology.xml#vowelharmony"/>
                <term ref="terminology.xml#atr"/>
        </index>
</div>
```

6.5.2.2 Descriptive prose

The majority of content in a grammar entry is plain text, which has been referred to as "descriptive prose" (see Section 6.3.2.1). Descriptive prose is structured into paragraphs, which is also the unit used in a TEI format. The element representing paragraphs is <p>. Paragraphs contain mostly text but also other smaller elements, which will be discussed in Section 6.5.3. Listing 6.7 shows a grammar entry with two paragraphs that do not contain any smaller elements.

Listing 6.7: Paragraphs

```
<div xml:id="ch2-sc4-4" type="section" n="2.4.4">
        <head>Vowel harmony</head>

        <p> Vowel harmony is a special type of assimilation where vowels are
        classified into two or more groups based on specific features and words
        usually consist of only vowels of one of those groups. Vowels in affixes
        then assimilate to the root vowels' group in order to fit into that
        group. </p>

        <p> In Ekiromi only some restrictions exist on which vowels can be in one
        word. In order to investigate vowel distribution, the sample contains
        only V.CV-words that are not borrowed. This is the only word structure
        where compounding can completely be precluded. The total number of items
        used here is 107.</p>
</div>
```

6.5.2.3 Exemplars

Exemplars have been defined as "embedded examples" (Section 6.3.2.2). As examples can be stored in a separate database, exemplars in the main grammar document are merely pointers to an example in the database. TEI-Grammar uses join groups <joinGrp> to represent exemplars. In order to be able to reference the exemplar elsewhere in the text, they need to carry an @xml:id attribute. Each <joinGrp> contains one or more pointer elements (<ptr>) to refer to the examples. The identifier of the example is therefore saved in the @target attribute of the <ptr> element. Listing 6.8 displays the structure of an exemplar in TEI-Grammar. The structure of examples within the example database will be discussed in Section 6.5.5.3. If the separation of an example database is not supported, the <annotationBlock> representing an example in the database can also be used directly in an entry.

Listing 6.8: The exemplar

```
<joinGrp xml:id="ex2-2" n="2.2">
        <ptr n="a" target="examples.xml#ibe057-00.009"/>
```

```
        <ptr n="b" target="examples.xml#ibe281-00.198"/>
        <ptr n="c" target="examples.xml#ibe327-00.069"/>
</joinGrp>
```

This method pertains to interlinear examples (textual examples that are not embedded in the prose). Lexical and textual examples embedded in descriptive prose can be encoded as linguistic units (Section 6.5.3.3) and lexical examples not embedded in the prose can be encoded as linguistic units in tables or lists (Section 6.5.2.4).

6.5.2.4 Tables, lists and figures

Structured description contains various forms of non-prose descriptions. Besides exemplars (Section 6.5.2.3), TEI-Grammar also aims at covering three more generic structures: tables, lists, and figures. Each of these has an established formalization within TEI. Other structures and formalisms may still be incorporated as figures in a picture format.

Lists are represented by the `<list>` element with nested `<item>` elements. They can also contain the `<head>` element to store a title. Like all grammar bricks, lists should have a unique identifier in the `@xml:id` attribute to allow referencing.

Listing 6.9: The list

```
<list xml:id="list2-1" n="2.1">
        <head>Suffix vowel harmony</head>
        <item>The object/possessive pronouns
                <foreign>na/no</foreign>
                <gloss>1SG.POSS/OBJ</gloss>,
                <foreign>sa/so</foreign>
                <gloss>2SG.POSS/OBJ</gloss>, and
                <foreign>ba/bo</foreign>
                <gloss>3PL.POSS/OBJ</gloss>
        </item>
        <item>
                The gerund <foreign>i- -aǹ </foreign> and
                <foreign>i- -oǹ</foreign>
        </item>
        <item>
                The pluractionalis forms <foreign>-adiàg</foreign>,
                <foreign>-odiòg</foreign>,
                <foreign>-àd</foreign>,
                <foreign>-òd</foreign>,
                <foreign>-àg</foreign>, and
                <foreign>-òg</foreign>
        </item>
</list>
```

Figures are represented by `<figure>` elements and also should have a unique identifier in their `@xml:id` attribute. They can contain a `<head>` element bearing the title and an empty `<graphic>` element which indicates the location of a picture in its `@url` attribute.

Listing 6.10: The figure

```
<figure xml:id="fig1-1" n="1.1">
        <head>The Abesabesi speaking communities</head>
        <graphic url="resources/images/Abesabesi-map.png"/>
</figure>
```

Tables are represented in TEI by the `<table>` element. They may contain a `<head>` element to store the title and multiple `<row>` elements to represent table rows. Within the rows, `<cell>` elements contain the text of each table cell.

Listing 6.11: The table

```
<table xml:id="tb2-3" n="2.3">
        <head>Oral vowels</head>
        <row>
                <cell/>
                <cell>front</cell>
                <cell>central</cell>
                <cell>back</cell>
        </row>
        <row>
                <cell>high</cell>
                <cell>i</cell>
                <cell/>
                <cell>u</cell>
        </row>
        <row>
                <cell>mid</cell>
                <cell>e ɛ</cell>
                <cell/>
                <cell>o ɔ</cell>
        </row>
        <row>
                <cell>low</cell>
                <cell/>
                <cell>a</cell>
                <cell/>
        </row>
</table>
```

	front	central	back
high	i		u
mid	e ɛ		o ɔ
low		a	

Table 6.3: Formatted version of Listing 6.11

Listing 6.11 displays the structure of a table containing Abesabesi's vowel inventory. A formatted version of the table can be seen in Table 6.3.

This is the classic encoding of a table in TEI. However, an application extracting data from such a table would not be able to recognize what kinds of units are stored in this table and how the units are distinguished from each other. In tables with linguistic units, the semantic markup discussed in Section 6.5.3.3 can be used to increase the machine-readability of tables. The phonemes in Listing 6.11, for example, could be marked as "phonemes". The column and row heads could be marked as "terms", as they describe phonological categories. Listing 6.12 displays a paradigm table of the Abesabesi independent pronouns (Table 6.4) that has been encoded with semantic markup (Section 6.5.3.3).

	SG	PL
1PERS	ònì	àbès
2PERS	òsì	àbèn
3PERS.HUM	àu	àba
3PERS.NHUM	ɛ̀jì	ìjì

Table 6.4: Independent pronouns

Listing 6.12: A paradigm table with semantic markup

```
<table xml:id="tb4-2" n="4.2">
    <head>
        <term ref="terminology.xml#idp">Independent pronouns</term>
    </head>
    <row>
        <cell/>
        <cell><term ref="terminology.xml#sg">Singular</term></cell>
        <cell><term ref="terminology.xml#pl">Plural</term></cell>
    </row>
    <row>
        <cell><abbr corresp="glossary.xml#firstP"/></cell>
        <cell>
            <w xml:lang="ibe" corresp="dictionary.xml#oni-2e12455">
```

```
                    ònì
                </w>
            </cell>
            <cell>
                <w xml:lang="ibe" corresp="dictionary.xml#abes-2e12342">
                    àbès
                </w>
            </cell>
        </row>
        <row>
            <cell><abbr corresp="glossary#secondP"/></cell>
            <cell>
                <w xml:lang="ibe" corresp="dictionary.xml#osi-2e15775">
                    òsì
                </w>
            </cell>
            <cell>
                <w xml:lang="ibe" corresp="dictionary.xml#aben-2e16664">
                    àbèn
                </w>
            </cell>
        </row>
        <row>
            <cell>
                <abbr corresp="glossary.xml#thirdP"/>.
                <abbr corresp="glossary.xml#hum"/></cell>
            <cell>
                <w xml:lang="ibe" corresp="dictionary.xml#au-2f20455">
                    àu
                </w>
            </cell>
            <cell>
                <w xml:lang="ibe" corresp="dictionary.xml#aba-2e23455">
                    àba
                </w>
            </cell>
        </row>
        <row>
            <cell>
                <abbr corresp="glossary.xml#thirdP"/>.
                <abbr corresp="glossary.xml#nhum"/>
```

```
            </cell>
            <cell>
                <w xml:lang="ibe" corresp="dictionary.xml#ɛji-2e17755">
                    èjì
                </w>
            </cell>
            <cell>
                <w xml:lang="ibe" corresp="dictionary.xml#iji-2e99955">
                    ìjì
                </w>
            </cell>
        </row>
</table>
```

The use of semantic markup in tables enables an application to recognize which linguistic units are used in the table cells and which dictionary entries or terminological concepts they are linked to. The relations between the different cells, however, cannot be extracted by an application. For example, it would not recognize that a phoneme in the phonemic inventory has the features indicated in its column and row head (e.g., /p/ is a voiceless bilabial plosive). To achieve this, a more complex encoding of grammar tables would need to be developed, which could possibly benefit from semantic web technologies. As tables in grammars are very diverse, this undertaking will not be part of this project. Along with a more complex encoding, paradigm tables and phonemic inventories could also be stored in a paradigm database to let readers access structured information directly.

6.5.3 Nanostructure

The nanostructure encompasses all elements that can be found within the grammar bricks: text, terms, abbreviations, linguistic units, references, and multimedia. While text is simply represented as text within elements and multimedia is represented as URL links to the resources (such as in a figure, Section 6.5.2.4), the other units of the nanostructure are discussed in the following subsections.

6.5.3.1 Terms

As discussed in Section 6.3.3.2, terms are structured in ontologies. TEI-Grammar stores these ontologies in the terminology database. Terms stored in the database can be used within the grammar bricks or as keywords for indexing the grammar entries (Section 6.5.2.1) or the examples (Section 6.5.5.3).

If a term is used in the descriptive prose, it can be tagged as `<term>`. This way, the term can be linked to an entry in the terminology database, where its affiliation to an ontology and the explanation of the term can be found. The attribute `@ref` carries the

identifier of the term in the ontology. In the example GET, readers can click on the tagged terms, which will take them to the index, where a brief definition is given as well as links to all instances of the term used in the text.

Listing 6.13: Terms

```
<p>
    Only the
    <term ref="terminology.xml#emphPro">emphatic pronoun</term>
    can occur in that position.
</p>
```

As mentioned by Nordhoff and Hammarström (2014, p. 172), terms could also be seen as named entities because they are names that link to a defined concept. If the logic of a semantic web is followed, other named entities, such as place or language names can be encoded using the `name` element. What is important is to use the attributes `@type` for the type of name (person, place, etc.) and `@ref` to identify an entity using a URI (ISBNs for books, ISO639-3 for languages, etc.). The encoding of named entities in TEI is discussed in Chapter 13 of the TEI Guidelines (TEI Consortium, 2020).

6.5.3.2 Abbreviations

Abbreviations can be used in any kind of text, such as in paragraphs, tables, or examples. They are similar to terms, as they are used in the text and lead to a concept within a structured list. Abbreviations are stored together with their expanded equivalents in an external abbreviation database, where they are grouped into a list (Section 6.5.4.4). Within the text, they are tagged using the `<abbr>` element with the `@corresp` attribute, which contains the identifier of the abbreviation in the abbreviations database. Listing 6.14 shows a paragraph with the abbreviation “VH” for vowel harmony. In this example, the element is empty, as the GET produces the abbreviation by drawing it from the abbreviations list through the identifier in the `@corresp` attribute.

Listing 6.14: Abbreviations

```
<p>
    Abesabesi attests a rather complex
    <abbr corresp="glossary.xml#vh"/> system.
</p>
```

Abbreviations in interlinear glosses of examples are very common in linguistic descriptions. When it occurs in the gloss line of an example, it labels a grammatical category and is usually formated in small capitals, such as in Example 6.3. This special formatting does not need to be indicated, as it can be inferred from the environment (being inside the gloss line of an example).

(6.3) *ɔ́ tu H ìwùd iʃo*
3SG be LOC inside house
's/he is inside the house' (ibe121-00.286)

6.5.3.3 Semantic markup: linguistic units

Grammar entries have a high frequency of words in other languages, especially words in the target language. Moreover, other linguistic units such as phones, phonemes, graphemes, and grammatical categories occur in grammatical descriptions. There are two reasons to represent them in a digital data format. Firstly, a distinction of linguistic units in the format also enables the application to distinguish these units. Detailed search queries (e.g., "Show all morphemes that are mentioned in a section with the index 'negation'!") are only possible if linguistic units are distinguished in the format. Secondly, it is not always evident for the readers that a token is a word of the target language or what linguistic unit the token is. In paper grammars, the different units are usually distinguished by formatting. Phonemes are written between slashes (e.g., /p/), phones in square brackets (e.g., [p]), and words or morphemes in the target language are italicized. Different formatting in a GET or even functionalities associated with a specific unit are only possible if the units are distinguished within the format.

Below, three possible levels for the encoding of linguistic units with TEI will be presented. Which level should be used depends on the author's available time and aspiration.

First level The first level has no distinction of linguistic units. All tokens in the target language and other linguistic units will simply be written into the `<p>` element of the descriptive prose or into other elements, such as tables or lists. The author might emulate the formatting from paper grammars by using slashes for phonemes. Italics could be encoded in TEI by the element `hi` and the value "italic" in the attribute `@rend`: `<hi rend="italic">àbès</hi>`. This level is not recommendable, as it misses the aim of TEI's semantic markup. The markup of the format should not be used to define the formatting but to specify the semantics of an element. Its formatting can then be derived from the semantics.

Second level A first step to a deeper semantic markup is to mark all tokens in the target language or in languages other than the metalanguage. In doing so, different linguistic units are not distinguished and all tokens in a language other than the metalanguage are marked equally. TEI provides the element `<foreign>` to mark words that do not belong to the default language of the document. `<foreign>` could be used for all kinds of linguistic units in a language other than the metalanguage: morphemes, lexemes, word forms, and even more complex units such as phrases, clauses, or sentences. It does not, however, differentiate them. The language can be indicated through the `@xml:lang` attribute. ISO 639-3 codes are recommended as identifiers for a specific language. Abesabesi, for example, has the code *ibe*. Listing 6.15 displays a paragraph of the Abesabesi sketch grammar

with two words not in the metalanguage, English. The first word *íni* is Abesabesi and the second word *àti* is a word from another language, Yoruba.

Listing 6.15: Foreign tokens

```
<p>
        Besides the fronted focus position, there are other contexts where
        independent pronouns must be used instead of bound pronouns.
        Firstly, they are required with the conjunction
        <foreign xml:lang="ibe">íni</foreign>
        or its Yoruba equivalent
        <foreign xml:lang="yor">àti</foreign>.
</p>
```

This simple markup already allows for an automatic distinction between target and metalanguage and for specific search queries that only apply to the target language. Furthermore, all tokens in the target language could be formatted in a uniform way.

Tokens in other languages are often used together with their meanings in what are known as "word-gloss" or "morpheme-gloss" pairs. The meaning of a foreign word, the "gloss", can be stored in the TEI element `<gloss>`. Listing 6.16 shows an example with two word-gloss pairs. The `<gloss>` element contains several abbreviations.

Listing 6.16: Glosses

```
<p>
        The forms <foreign xml:lang="ibe">àbès</foreign>
        <gloss><abbr corresp="glossary.xml#1pl"/>
        <abbr corresp="glossary.xml#idp"/></gloss>
        and <foreign xml:lang="ibe">àbèn</foreign>
        <gloss><abbr corresp="glossary.xml#2pl"/>.
        <abbr corresp="glossary.xml#idp"/></gloss>
        are identical with the long forms of the
        <abbr corresp="glossary.xml#1pl"/> and
        <abbr corresp="glossary.xml#2pl"/>
        possessive and object pronouns.
</p>
```

Through this encoding, text in the target language and glosses can be marked and distinguished from plain text. This could be used by the GET to format elements differently. While plain text does not receive special formatting, text in the target language is rendered in italics and glosses, between quotation marks. For instance, the formatted output of Listing 6.16 could look like this: "The forms *àbès* '1PL.IDP' and *àbèn* '2PL.IDP' are identical with the long forms of the 1 and 2PL possessive and object pronouns."

However, this encoding does not indicate which gloss is attributed to which token. As tokens in the target language, like words or morphemes, could be seen as references

to a lexical entry, a link could be added. The `<foreign>` element can take a `@corresp` attribute to establish a link to the lexical entry in the lexical database. A gloss likewise refers to a lexical entry, as it is the gloss for a specific sense of the lexical entry. Therefore, `<gloss>` elements can also take a `@corresp` attribute to store a link to the lexical entry. An application can now detect that a token in the target language belongs to a specific gloss.

Listing 6.17: Links to lexical entries

```
<foreign xml:lang="ibe" corresp="dictionary.xml#abes-2e12342">àbès</foreign>
<gloss corresp="dictionary.xml#abes-2e12342">
        <abbr corresp="glossary.xml#1pl"/>
        <abbr corresp="glossary.xml#idp"/>
</gloss>
```

In some cases, text in the target language and the associated gloss cannot be linked to a lexical entry, as they contain a more complex unit such as a phrase or sentence. In this case, glosses could use the `@corresp` attribute to create a link to the `<foreign>` element. Therefore, the `<foreign>` element needs an identifier (Listing 6.18).

Listing 6.18: Links between clause and translation

```
<foreign xml:lang="ibe" xml:id="simpleClause1">ɛŋɔ hu</foreign>
<gloss corresp="#simpleClause1">The snake died.</gloss>
```

Third level The third level involves a markup that distinguishes different linguistic units. As deep semantic markup is rather time-consuming, it should be optional. TEI offers a range of elements for linguistic units in the module, “Analysis and Interpretation”. The following list displays units that could be differentiated in a grammatical description.

- phonological units
  - (allo-)phone
  - phoneme
  - grapheme
  - syllable
- form units
  - (allo-)morph
  - morpheme
  - lexeme
  - word form
  - construction

  - intonation contour
  - phrase
  - clause
  - sentence

- function units
  - meaning
  - grammatical categories

TEI only has some element tags to represent these units, such as `<w>` for “word” or `<cl>` for “clause”, but other units cannot be represented through a specific element tag. However, there is the arbitrary segment element (`<seg>`) that can be used for all segments and specified through the `@type` attribute. In order to have a uniform layout and to avoid representing some units through a specific tag and others through the arbitrary segment tag, all linguistic units should be represented through the arbitrary segment tag and be specified through the `@type` attribute.

Phonological units are all units relevant in the grammar entries that deal with phonetics, phonology, or orthography. Besides the atomic units, “phoneme”, “phone”, and “grapheme”, they also include more complex units like the “syllable” or other metric units.

Listing 6.19: Glosses

```
<p>
    Closed syllables such as <seg type="syllable">ton`</seg> or
    <seg type="syllable">lam</seg> only occur with the nasal
    consonants <seg type="phoneme">n</seg>,
    <seg type="phoneme">m</seg>, or
    <seg type="phoneme">ŋ</seg> as coda. Note, that
    <seg type="phoneme">n</seg> assimilates with succeeding
    consonants and can also be realized as
    <seg type="phone">m</seg> and <seg type="phone">ŋ</seg>.
    Despite the assimilation, the phoneme is represented as
    <seg type="grapheme">n</seg> in the orthography.
</p>
```

Morphosyntactical units have a form and a function part. As described by Nordhoff (2012b), most grammar entries either describe such a unit from form to function or from function to form. In the first case, a formal unit is the “lemma” and the form is connected to one or more functions. In the other case, a function unit is the “lemma”, which is connected to one or more forms. Form units and function units are also mentioned separately.

Such a form unit can be a “morpheme”, an “intonation contour”, or a “construction”, as listed by Nordhoff (2012b, p. 44). Describing the form of a morpheme also includes its different expressions. Therefore, “(allo-)morphs” are also formal units that occur in grammatical descriptions. Furthermore, the use of morphemes or intonation contours may be illustrated by attaching them to words. In this way, “word forms” are also formal units to be included. Another unit that occurs frequently in grammatical descriptions is the “lexeme”. Lexemes are references to a specific form-function entry in the dictionary by mentioning the lemma. Sometimes, the lemma is mentioned together with the meaning. This combination has been referred to as “word-gloss pair”. Lexemes or word-gloss pairs may be used in grammatical descriptions for various reasons. They might be part of a construction; they might be an element that imposes a constraint on a certain morphosyntactic unit; they might be used as lexical examples for other phenomena; or they might be used in the descriptive prose to explain a textual example. More complex examples might be phrases, clauses, sentences, or utterances. While it is recommended for these units to be presented as IGT, they also occur within the descriptive prose – often accompanied by a translation.

TEI-Grammar uses the element `<seg>` with the `@type` attribute `"morpheme"` or `"morph"` instead of the element `<m>`. If morphemes are stored in the lexical database or even in a separate morpheme database, the entry can be linked using the `@corresp` attribute. The `@subtype` can be used to indicate the type of morpheme (stem, prefix, etc.). Listing 6.20 shows a paragraph where the morpheme *kV-* and its allomorphs *ka-*, *ke-*, and *ko-* are encoded using the element `<m>`. All elements are linked to the same lexical entry.

Listing 6.20: Glosses

```
<p>
        The prefix <seg type="morpheme" xml:lang="ibe"
        corresp="dictionary.xml#kV-pre-19982" subtype="prefix">
        kV-</seg> denotes the perfect tense. Its vowel
        harmonizes with the verb's <abbr corresp="glossary.xml#atr"/>
        value. Resulting allomorphs are
        <seg type="morph" xml:lang="ibe" corresp="dictionary.xml#kV-pre-19982"
        subtype="prefix">ka-</seg>,
        <seg type="morph" xml:lang="ibe" corresp="dictionary.xml#kV-pre-19982"
        subtype="prefix">ke-</seg>, and
        <seg type="morph" xml:lang="ibe" corresp="dictionary.xml#kV-pre-19982"
        subtype="prefix">ko-</seg>.
</p>
```

Word forms and lexemes can be indicated through the `@type` attributes `"word"` and `"lexeme"`. The reference to a dictionary entry can be encoded using `@corresp`. This way, a word’s base form can be retrieved from the dictionary entry. Words that contain more than one morpheme can be further segmented into morphemes using the `<seg>` element. This method can also be used for all other complex units.

More complex units can be encoded using the types `"phrase"`, `"clause"`, `"sentence"`, and `"utterance"`. Likewise, an intonation contour can be encoded using the type `"contour"`. Moreover, a link to a lexical entry could be added through the `@corresp` attribute if intonation contours are stored in the lexical database. The HLH contour for conditional clauses in Abesabesi, for example, could be encoded as follows:

```
<seg type="contour" corresp="dictionary.xml\#HLH-2342">HLH</seg>
```

Intonation contours could also benefit from a `@baseForm` attribute, since they can have different realizations. A tonal melody HL could, for example, be realized HLL on a bisyllabic word and HLLL on a trisyllabic word. For a `<seg>` element, however, the `@baseForm` attribute is not permitted in TEI.

Constructions are a combination of lexemes, morphemes, place holder units, and rules or constraints. Some grammatical descriptions, including the Abesabesi Grammar Sketch (Chapter 3), do not attempt a formalization but merely describe the constructions. Others, however, use semi-standardized formalizations and additionally describe possible rules or constraints that could not be captured in the formalization. As constructions are formalizations that heavily rely on a particular linguistic theory, no attempt to encode its inner structure will be presented. I recommend using only the `<seg>` element with `"construction"` as the `@type` value and plain text within the element. They do not necessarily correspond with other complex units such as a phrase, clause, or sentence, which is why none of those TEI elements can be used. Usually, they are displayed as a mix of lexemes, morphemes, and variables. For more complex formalizations, picture files could be used. Another method is to avoid formalizations and only rely on descriptions. A label can be chosen for each construction to be able to refer to it elsewhere in the grammar. The following example uses the label “negative focus construction”. The `@corresp` is used to refer to the grammar entry, where the construction is described.

```
<seg type="construction" corresp="grammar.xml#ch9-sc4.5">
        negative focus construction
</seg>
```

All of these form units, except for constructions where a label in the metalanguage is used, require a language attribute (`xml:lang`) to differentiate the target language from the metalanguage. If a form unit is further segmented using other elements, only the highest element needs to have the language attribute.

Form units can be used as lexical examples but not all form units are necessarily lexical examples. Form units can, for example, also be used within a grammar entry because the entry deals with the form and function of said unit. In this case, the unit cannot be referred to as an example.

A function unit contains the meaning that is related to one or more form units. It cannot be encoded using the arbitrary segment element, as it is not a segment. While lexemes have a meaning that is stored in a dictionary, grammatical form units have a grammatical function. A grammatical function can be described by one or multiple grammatical

categories. The function of the Abesabesi prefix *kV*, for example, can be described with the grammatical category "perfect" or its abbreviation PRF. The function of the pronoun *ìjì*, on the other hand, needs to be described with multiple grammatical categories: "third person", "plural", "non-human", and "independent pronoun" (3SG.NHUM.IDP).

Two types of grammatical categories need to be distinguished: universal categories and language-specific categories. A distinction between these two is often neglected in grammar writing, although it is essential to avoid confusion. Haspelmath (1993)'s Grammar of Lezgian is a counterexample where language-specific categories are capitalized in order to differentiate them from universal categories. In a digital format for reference grammars, this distinction is achievable. As the terminology database can distinguish multiple ontologies, two separate ontologies can be established for universal and language-specific categories. Grammatical categories can be treated as terms, as they can be defined in an ontology and tend to be influenced by a theory, such as BLT. They occur in descriptive prose as table heads in paradigm tables or within a gloss of a word (either in interlinear examples or in a word/morpheme-gloss pair) and may be encoded using the element `<term>` (Section 6.5.3.1). As abbreviations also refer to a term, the element `<abbr>` can be used likewise.

TEI mentions the encoding of grammatical categories only in the dictionary module. The element `<gram>` can be used to represent a grammatical category as a key-value pair. The category "singular" is, for example, a value of the category "number". An abbreviation or term could be placed into such a `<gram>` environment in order to specify the key-category. `<gram>` elements can be grouped into groups (`<gramGrp>`). For example, the grammatical categories for the Abesabesi pronoun *ìjì* could be encoded as follows:

```
<gramGrp>
        <gram type="person"><abbr corresp="glossary.xml#3"/></gram>
        <gram type="number"><abbr corresp="glossary.xml#pl"/></gram>
        <gram type="animacy"><abbr corresp="glossary.xml#nhum"/></gram>
        <gram type="pos"><abbr corresp="glossary.xml#idp"/></gram>
</gramGrp>
```

In the example, the `@type` attribute has been used to indicate the key and the text within the element has been used to indicate the value of the grammatical category. The use of `<gram>` and `<gramGrp>`, however, is restricted to dictionary entries. It is not permitted within the `<gloss>` element, where it would be required to mark the function of a form unit.

As mentioned in Level 2, formal units often come in pairs with their corresponding functional unit. These pairs are sometimes called "word-gloss" or "morpheme-gloss" pairs but could also apply to other form units that bear meaning. This level attempts to encode the pairs as one unit with two parts. Segmental elements in TEI (such as `<m>` or `<w>`) are designed to only contain the form part of a segment. From a linguistic perspective, it should contain both the form and the function. To use the element `<entry>` from TEI's dictionary module instead of these segmental elements would not be suitable, as linguistic

units in grammatical descriptions are not entries where all kinds of information about a segmental unit is stored. Instead, they are merely references to the entry or instances of the entry. It would be desirable for all segmental elements in TEI to permit the elements `<form>` and `<sense>` within. This is currently not possible in TEI but would allow for a structure of word-gloss pairs like the one in Listing 6.21

Listing 6.21: Word-gloss pair

```
<p>
        The forms
        <w xml:lang="ibe" lemmaRef="dictionary.xml#abes-2e12342">
                <form><orth>àbès</orth></form>
                <gloss>
                        <gramGrp>
                                <gram type="person">
                                        <abbr corresp="glossary.xml#1"/>
                                </gram>
                                <gram type="number">
                                        <abbr corresp="glossary.xml#pl"/>
                                </gram><pc>.</pc>
                                <gram type="pos">
                                        <abbr corresp="glossary.xml#idp"/>
                                </gram>
                        </gramGrp>
                </gloss>
        </w>and
        <w xml:lang="ibe" lemmaRef="dictionary.xml#aben-2e12309">
                <form><orth>àbèn</orth></form>
                <gloss>
                        <gramGrp>
                                <gram type="person">
                                        <abbr corresp="glossary.xml#2"/>
                                </gram>
                                <gram type="number">
                                        <abbr corresp="glossary.xml#pl"/>
                                </gram><pc>.</pc>
                                <gram type="pos">
                                        <abbr corresp="glossary.xml#idp"/>
                                </gram>
                        </gramGrp>
                </gloss>
        </w> are also used as possessive and object pronouns.
</p>
```

This structure containing the form and function of a unit is only possible if the gloss immediately follows the form within the prose. There are, however, also sentences where form and gloss are separated – for example: "If used after nouns, the pronoun 'àbès' functions as a possessive pronoun and has the meaning '1PL.POSS'." In this case, the gloss has to be encoded separately. As mentioned in Level 2, it can be linked to the lexical entry or the form unit.

6.5.3.4 References

As discussed in Section 6.3.3.4, there are five types of references:

- References to lexical entries
- References to corpus data
- References to other elements within the grammar
- References to other publications
- References to terms drawn from ontologies

Linguistic units within the prose, tables, or interlinear examples can refer to lexical entries (Section 6.5.3.3). The `<foreign>`, `<seg>`, or `<gloss>` elements used for linguistic units can each contain a `@corresp` attribute, where links to lexical entries in the lexical database can be established.

```
<seg type="lexeme" corresp="dictionary.xml#aje_23b232">áje</seg>
<gloss corresp="dictionary.xml#aje_23b232">mother</gloss>
```

Corpus data can be referenced from textual examples within the prose or as interlinear examples. If utterances or other complex linguistic units have their source in corpus material, this should be indicated in order to enable readers to verify any claims. Corpus data is represented in the text database, where transcribed utterances of the original recordings can be saved. Data within the text database itself is connected to the corpus data (Section 6.5.5.2). Textual examples only need to refer to the source utterance within the text database. Textual examples within the prose are treated as complex linguistic units and can use the `@corresp` attribute to refer to the transcribed utterance. Interlinear examples in the example database also use the `@corresp` attribute to refer to the utterance in the text database. This is explained in Section 6.5.5.2 and illustrated in Figure 6.10.

References to other elements within the grammar (other grammar entries, exemplars, figures, tables, or lists) are particularly common in paper grammars and all make use of a unique identifier, such as numbers, to reference a specific entity. In TEI, we can use the `<ref>` (reference) or `<ptr>` (pointer) element with a `@target` attribute to reference an entity within or outside the document. In the Abesabesi grammar document, all elements that are within the document and have a unique identifier (`@xml:id`) can be referenced through a `@target` attribute starting with a hash (#) and followed by the identifier.

The section `<div xml:id="ch1-sc4"/>`, for example, can be referenced as follows: `<ref target="#ch1-sc4"/>`.

References to elements outside the main grammar document are similar. Websites can be referenced by simply putting the URL into the `@target` attribute (Listing 6.22).

Listing 6.22: References to websites

```
<p>
    The <ref target="http://hdl.handle.net/2196/a6371e17-d083-4e29-bc4a-
    ecb9303f9197">Abesabesi Collection</ref>
    contains all data this grammar sketch is based on.
</p>
```

Elements form other documents, such as publications from the bibliography database or terms from the terminology database, can be referenced by the document name, the hash (#), and the element identifier. As terms have their own reference attribute `@ref`, they do not require an additional `<ref>` element (Section 6.5.3.1):

`<term ref="terminology.xml#noun-classification"/>`.

If the `<ref>` element does not contain text, the `<ptr>` should be used instead to generate the section, example, or table number. This number can be retrieved from the `@n` attribute of a referenced element. For instance, the paragraph encoded in Listing 6.23 can be rendered as “The process of complex nominalization is discussed in Section 7.2.2.”

Listing 6.23: The use of pointers

```
<p>
    The process of complex nominalization is discussed in Section
    <ptr target="#ch7-sc2-2">.
</p>
```

Likewise, pointers to a resource within the bibliography database can be rendered in any citation style by using data from the bibliography entry. The encoded reference in Listing 6.24 could thus be rendered as “The four dialects of Abesabesi are mainly based on vowel harmony types (Agoyi, 2008)”.

Listing 6.24: References to publications

```
<p>
    The four dialects of Abesabesi are mainly based on vowel harmony types
    <ptr target="bibliography.xml#agoyi2008" />
</p>
```

6.5.4 Databases

Besides the main grammar document, a couple of databases are used to save specific kinds of data: the example database, the text database, the dictionary database, the

bibliography database, the terminology database, and the abbreviations database. All of these six databases can be designed as TEI documents. Since the example database and the text database require a method to encode IGT which has not been covered by TEI before, they are discussed in a separate section (Section 6.5.5)

6.5.4.1 Lexical database

Lexical entries are stored in the lexical database. The TEI P5 Guidelines (TEI Consortium, 2020) discusses the structure of dictionaries in Chapter 9. Any valid TEI-dictionary could be referenced from the grammar file. The entries only need to have unique identifiers to allow for referencing. As the encoding of lexical data in dictionaries has been extensively discussed in Chapter 9 of the TEI guidelines and has its own module, it will not be discussed further in this thesis.

The lexical database of the Abesabesi Sketch Grammar is based on a dictionary compiled by the software, FLEx. The output produced by FLEx was transformed into TEI. Listing 6.25 displays the transformed dictionary entry of the verb *ìjag* 'ocra seed' in order to exemplify an entry of the lexical database. It stores the form in the element `<form>` and the sense in the element `<sense>`. The element `<gramGrp>` stores grammatical information such as the part of speech information (`<gram type="pos">`) and information about the VH type (`<gram type="vh">`).

Listing 6.25: The structure of a dictionary entry

```
<entry xml:lang="ibe" xml:id="ìjag_066d8bfd-459d-42a2-9e1c-91c6dbf4ed02">
    <form type="lemma">
        <pron notation="ipa">ìjag</pron>
    </form>
    <sense>
        <gloss>okra_seed</gloss>
        <cit type="translationEquivalent">
            <quote xml:lang="en">
                okra seed
            </quote>
            <gramGrp>
                <gram type="pos">
                    <term ref="terminology.xml#noun"/>
                </gram>
                <gram type="vh">{Infl}[ATR(o):na]</gram>
            </gramGrp>
        </cit>
    </sense>
</entry>
```

As mentioned in Sections 4.4.1 and 5.1.3, consistent use of terminology within the grammatical description and also among all grammar documents is important to help the reader make connections and identify categories in different parts of the grammar. The `<term>` elements can be used to represent grammatical features in the `<gram>` elements. Terms used in the lexical entries can be linked to terms in the terminology database. It is currently not possible to link the `@type` of the `<gram>` elements to the terminology database. Recommendations on how that could theoretically be changed will be given in Section 6.5.6.

6.5.4.2 Bibliography database

The bibliography database is a list of all published resources cited within the grammar. Like dictionaries, the encoding of bibliographies has also been extensively discussed in Section 3.11.1 of the TEI P5 guidelines (TEI Consortium, 2020). As bibliographies of grammars are not structurally different from those of any other text, this method of encoding can also be used within TEI-Grammar. The bibliography used for the Abesabesi Sketch Grammar has been exported from a literature management software and also transformed to TEI. An example entry can be seen in Listing 6.26

Listing 6.26: The structure of a reference within the bibliography database

```
<biblStruct>
    <monogr>
        <author>
            <persName>
                <forename>Ben</forename>
                <surname>Elugbe</surname>
            </persName>
        </author>
        <title level="m">Comparative Edoid: phonology and lexicon</title>
        <imprint>
        <pubPlace>Port Harcourt</pubPlace>
        <publisher>University of Port Harcourt Press</publisher>
        <date when="1989"/>
        </imprint>
    </monogr>
</biblStruct>
```

6.5.4.3 Terminology database

The terminology database stores terms and their explanations and attributes them to different ontologies. As ontologies are structured like trees, they can be represented as nested lists in the `<body>` element of a TEI document. Each base-level list in the

terminology database thus represents one ontology. A `<head>` element can be attributed to a list in order to indicate the name of the ontology. Each term is represented within the list as an `<item>`. The item contains a `<term>` and a `<gloss>`, where a definition for the term is given. In order to represent the nested structure, items can contain another list of items. Each item needs to receive a unique identifier so that terms can be referenced from the main grammar document (Section 6.5.3.1). Listing 6.27 displays an example ontology with language-specific categories for Abesabesi.

Listing 6.27: Terms within the terminology database

```
<list type="ontology" xml:id="myOntology">
    <head>Abesabesi Categories</head>
    <item xml:id="noun">
        <term>Noun</term>
        <gloss>
            The noun is a word class that refers to a broad range of
            entities or experiences in the world. In contrast to verbs,
            they start with a vowel.
        </gloss>
        <list>
            <item xml:id="noun-classification">
                <term>Classification </term>
                <gloss>
                    Nominal classification is a phenomenon in
                    some languages where nouns are grouped in
                    noun classes according to semantic traits.
                    Noun classes can be morphologically marked.
                </gloss>
            </item>
            <item xml:id="noun-functionNoun">
                <term>Function Noun</term>
                <gloss>
                    Function nouns are a subclass of nouns.
                    They express a grammatical or semantic
                    relation with another constituent of the
                    clause and are thus semantically similar to
                    prepositions.
                </gloss>
            </item>
            <item xml:id="noun-property">
                <term>Property Noun</term>
                <gloss>
                    Property nouns are a subclass of nouns.
```

```
                                    They express a property and can be used
                                    alone or as nominal modifiers.
                                </gloss>
                            </item>
                        </list>
                    </item>
</list>
```

6.5.4.4 Abbreviation database

The abbreviation database has a similar structure to that of the terminology database. It also consists of lists of items in the `<body>` element of a TEI document. Again, each item requires a unique identifier in order for the term to be referenced in the main grammar document. Within the items, the element `<abbr>` can be used to store the abbreviation and `<expan>` to store the expanded term. However, the abbreviation database does not require a nested structure. Listing 6.28 displays a list of abbreviations with three items.

Listing 6.28: Abbreviations within the abbreviation database

```
<list>
        <item xml:id="fut">
                <abbr>FUT</abbr>
                <expan>future</expan>
        </item>
        <item xml:id="nfut">
                <abbr>NFUT</abbr>
                <expan>Non-future</expan>
        </item>
        <item xml:id="hab">
                <abbr>HAB</abbr>
                <expan>Habitual</expan>
        </item>
</list>
```

As abbreviations are in fact abbreviated terms and named entities, a DRG could also combine the terminology and abbreviation databases to one single database. Each item could then contain a term, an abbreviation, and a gloss. The `<expan>` element would not be required as the term already contains the expanded term. This connection between terms and abbreviations could also be represented in separate databases by adding a `@corresp` attribute to an abbreviation item to link it with an existing term in the terminology database.

6.5.5 Interlinear Glossed Text

Having discussed the use and structure of the IGT format in Section 6.3.2.2, the aim of this section is to present a method of storing IGT in a digital format, which will be referred to as "TEI-IGT". Within the DRG, TEI-IGT will be used to store texts in the text database and examples in the example database. As the aim of this project is to propose a comprehensive digital TEI format for the whole of a reference grammar, IGT will also be encoded in TEI.

Which functional requirements such a format needs to fulfill can best be derived from how texts and examples are processed and used within a DRG. In the model DRG proposed in Section 5.4, the main agent to process IGT is what I referred to as an IGT editor. While Section 5.4.3.4 only abstractly described its mode of operation, the following paragraphs will sketch out a more detailed workflow for processing IGT within the editor. This outline of the editor's processes will shed light on the requirements for the TEI-IGT.

Processing texts and examples to make them available in a grammatical description involves several stages. The model process mentioned in Section 6.3.2.2 involves the following stages from a recorded utterance to an exemplar in a grammar entry:

RECORDED UTTERANCE

→ TRANSCRIBED UTTERANCE

→ EXAMPLE (MODIFIED AND FORMATED UTTERANCE)

→ EXEMPLAR (EMBEDDED EXAMPLE)

The two databases containing IGT, the text database and the example database, store the products of the two intermediate stages. While the text database stores transcribed utterances, the example database stores examples. Recorded utterances, on the other hand, are ideally stored in an archive and exemplars are merely pointers within the grammatical description that refer to the respective example in the example database. The content and interaction between the grammar documents that deal with different example stages is modeled in Figure 6.10.

The proposed workflow for text and example processing in an IGT editor is illustrated in Figure 6.11. Transcribed utterances that are produced with transcription software, such as ELAN, can be imported by the editor. These utterances are converted into TEI-IGT and are equipped with links to their respective recording, its metadata, and the time frame within the recording. Transcribed utterances can also be incorporated manually in case no archive deposit exists. All of those utterances are stored in the text database. The editor would also be a management tool for the utterances. This means that utterances can be grouped into texts and texts can be selected to be published in the GET. As a default, utterances will be grouped into texts according to the recording they have been imported from. Texts are thus a collection of utterances in the IGT format. As they are

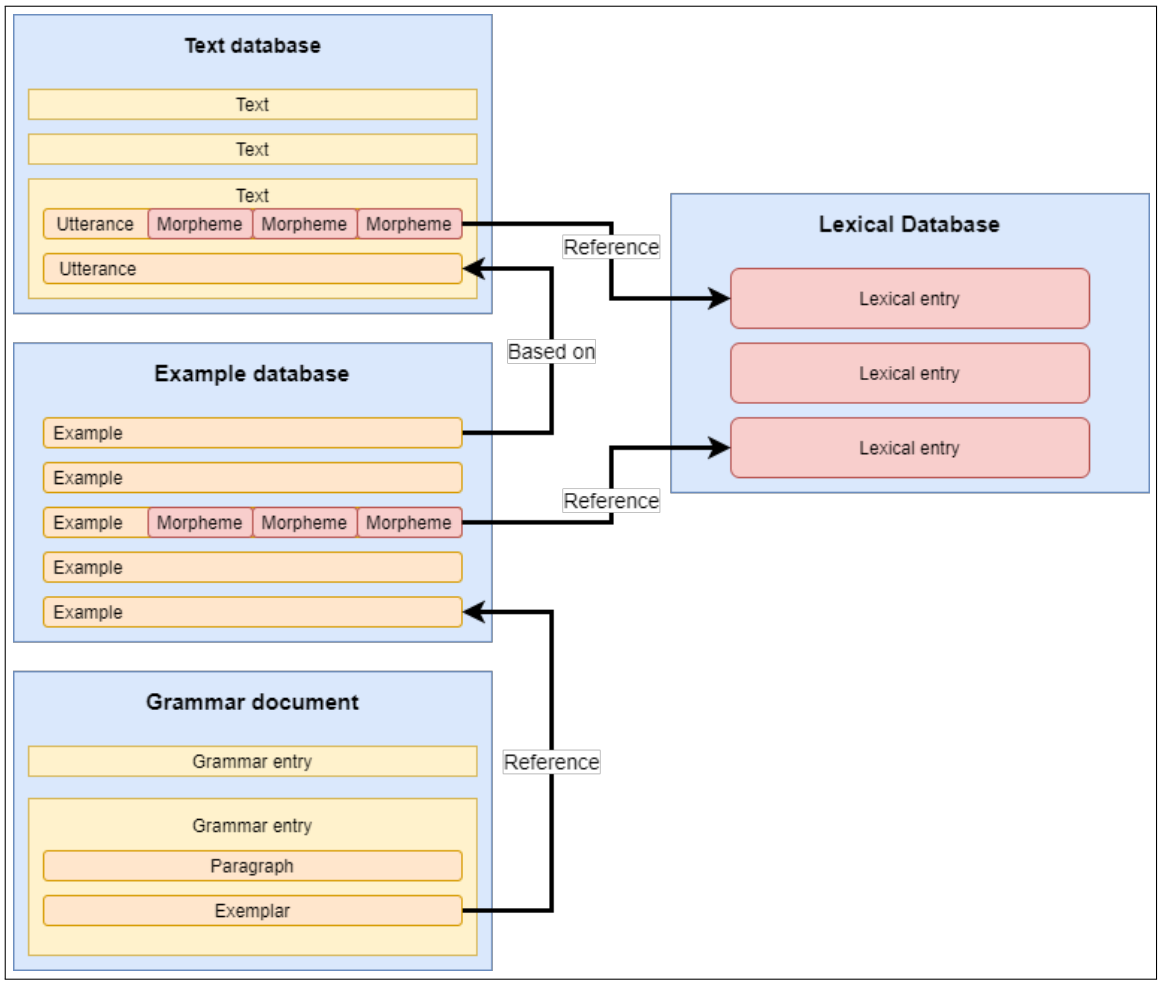


Figure 6.10: Interaction between example stages

coherent units, they should be formatted in a consistent way. Formatting choices should then be stored on the text level.

Examples are seen as modified and formatted utterances. As mentioned by Good (2004, Section 1.4), interlinear examples often deviate from the standard presentation to allow for specific illustration of a phenomenon. Formatting choices thus need to be saved on the utterance level. Content of the transcribed utterances in the text database is copied into the example database and refers back to the transcribed utterance which, in turn, refers back to the recorded utterances. This three-way distinction is indispensable for separating recordings from original texts and from examples, where the form and sometimes even the content has been modified. Within the editor, the content of examples can be modified and formatting choices can be made. The author needs to have a say in which tiers are shown, how far the utterances are segmented, where glosses align, which elements should be highlighted, etc. Moreover, the editor should allow for manual ingestion of examples. If they are taken from original utterances, the transcribed utterance should be ingested into the text-database and then used as an example. If they

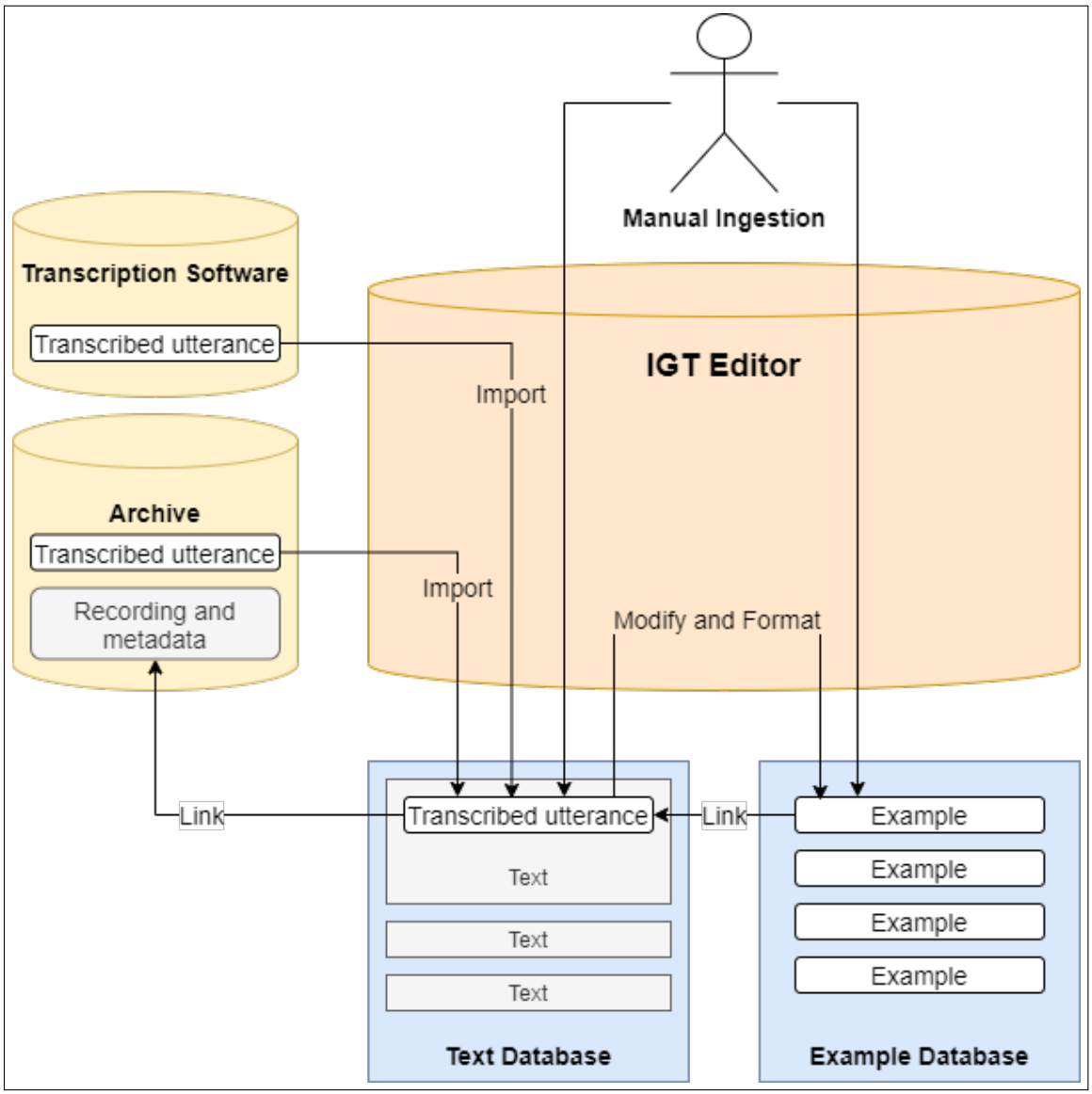


Figure 6.11: Complex workflow of the IGT editor

do not refer back to an original utterance (e.g. examples created by introspection), they can be added manually within the editor and directly stored as examples in the example database.

As most examples go through the phase of being a transcribed utterance, the text database not only contains those texts selected for text collection visible in the GET but also other utterances that might only be used as examples within the grammar.[9]

[9]In this workflow, I have deliberately disregarded the fact that texts usually also deviate from the original transcriptions as they might have been edited. This has been done to simplify the process and to avoid the need of a third IGT database for published texts. As the TEI-IGT can mark links to original utterances, however, this separation would be possible.

The workflow has pointed out four main functions of TEI-IGT:

- Store transcribed and annotated utterances in an IGT format
- Provide links from examples to transcribed utterances and from utterances to recordings and metadata
- Store the author's formatting choices for examples (and texts)
- Group utterances into texts

A couple of formats have already been proposed for the storage of IGT data: the Shoebox/Toolbox format developed by SIL[10], an XML format sometimes referred to as BHB after the developers (Bow, Hughes, & Bird, 2003), the IGT-XML by Palmer and Erk (2007), and, what seems to be the latest standard, the Extensible interlinear glossed text (Xigt) format by Goodman, Crowgey, Xia, and Bender (2015). I will leave the comparison and evaluation of those formats to Goodman et al. (2015, pp. 464-466) and concentrate on the latest standard, Xigt, on whose principles TEI-IGT will be based.

The Xigt format, as well as the other formats, was designed as a storage format for IGT. This is why formatting choices are not included in the data model. As it is a storage format, it can be rendered using a transformation processor and formatting depends on stylesheets. Examples in grammatical descriptions, however, cannot be separated from their formatting, as the formatting – such as the order of the tiers, highlighting elements, or indicating complex units through brackets – also transmits semantics. Therefore, examples in a grammar cannot all be rendered in the same way using a stylesheet. Each example needs to have the formatting choices of the author encoded.

The main purpose of Xigt is the automatic processing of IGT data, as it is done in the ODIN project (W. D. Lewis & Xia, 2010), where IGT data from linguistic documents on the web was processed and stored in a database[11].

Xigt's structure is rather flat. This is intended to enable fast processing and flexibility. Its root element, the `<xigt-corpus>` can contain several `<igt>` instances. These instances contain all segmentation and annotation tiers for one utterance. Each tier is represented by a `<tier>` element and contains `<item>` elements. An additional `<metadata>` element may be placed under `<xigt-corpus>`, `<igt>`, or `<tier>` to store metadata.

`<tier>` types are differentiated by their type attribute. Xigt provides a core of five basic tiers that can be extended by creating customary tier types and aligning them to the other ones. The five tier types are "phrases", "words", "morphemes", "glosses", and "translations".

Segmentation and alignment are represented by "alignment expressions". They mainly work with ID referencing as well as with slicing and joining strings of characters. This

[10] retrieved 2021-03-24, from `https://software.sil.org/toolbox/`

[11] A similar approach for PDF scans of grammars was conducted later by Round, Macklin-cordes, Ellison, and Beniamine (2020). The format used to store the results is less complex as it was designed for identification and extraction of IGT in PDF copies with high accuracy and precision.

way, Xigt can also represent one-to-many and many-to-one alignments. While parts of the text can be referenced from higher tiers to copy the content, Xigt makes use of "shadowing", where the content of a referenced item is overwritten if declared new in the depending tier. This can, for example, be used if phonological processes blur the lines of morphemes so that morphemes have a different structure in the morpheme tier than they have in the words tier.

Figure 6.29 displays an example serialization of Xigt data from Goodman et al. (2015, p. 474).

Listing 6.29: Xigt serialization

```
<?xml version="1.0" encoding="utf-8"?>
<xigt-corpus>
    <igt id="i1">
        <tier type="words" id="w">
            <item id="w1">cocinas</item>
        </tier>
        <tier type="morphemes" id="m" alignment="w">
            <item id="m1" alignment="w1[0:5]"/>
            <item id="m2" alignment="w1[5:7]"/>
        </tier>
        <tier type="glosses" id="g" alignment="m">
            <item id="g1" alignment="m1">cook</item>
            <item id="g2" alignment="m2">2</item>
            <item id="g3" alignment="m2">SG</item>
            <item id="g4" alignment="m2">PRS</item>
            <item id="g5" alignment="m2">IND</item>
        </tier>
        <tier type="translations" id="t" alignment="w">
            <item id="t1" alignment="w1">(You) cook</item>
        </tier>
    </igt>
</xigt-corpus>
```

6.5.5.1 TEI-IGT

TEI-IGT is based on the basic principles of Xigt while maintaining the logic of TEI. It also extends the functionalities of Xigt as it allows for encoding of formatting choices. In creating a TEI format for IGT, it will be possible to apply simple search queries across the grammatical description as well as the IGT databases.

The basic elements of the Xigt structure are represented by TEI elements as follows. The two databases for examples and texts are both TEI documents. IGT instances are directly nested inside the `<body>` element of the TEI document, which would be the equiva-

lent of the <xigt-corpus> element. <igt> elements are represented by <annotationBlock> elements. In contrast to Xigt's <tier> element, TEI-IGT has two classes of tiers: Segmentation tiers and annotation tiers. While segmentation tiers contain content of the target language in different depths of segmentation, annotation tiers annotate the whole utterance or the segmented pieces. Although Xigt does not make this differentiation, it is rather important for rendering and formatting the IGT instances. TEI-IGT differentiates them by using the utterance element (<u>) for segmentation tiers and the span group element (<spanGrp>) for annotation tiers. Each tier can have multiple items that Xigt represents as <item> elements. In TEI, segments of segmentation tiers are represented as arbitrary segments (<seg>) and annotation items in annotation tiers are represented as spans (). Goodman et al. (2015, p. 465) mentions that a single element type representing the content in each tier is required to define general processing rules for an arbitrary level of annotation. This cannot be followed because a separation between segmentation tiers and annotation tiers is important. However, a single element type is used within segmentation tiers and a single element type is used within annotation tiers. In order to differentiate the linguistic segmentation units or annotation items, the @type attribute should be used on <seg> and elements as described for linguistic segmentation units in Section 6.3.3.1.

Listing 6.30: Basic structure of TEI-IGT

```
<body>
    <annotationBlock>
        <!--segmentation tiers-->
        <u>
            <seg type="..."> ... </seg>
        </u>
        <!--annotation tiers-->
        <spanGrp>
            <span type="...">...</span>
        </spanGrp>
    </annotationBlock>
</body>
```

Tiers can, like in Xigt, have types, which are indicated in the @ana attribute. Listing 6.31 displays the five basic tier types that are used in most digital IGT standards: phrases, words, morphemes, glosses, and translation.

Listing 6.31: Tier types

```
<u ana="#phrases">...</u>
<u ana="#words">...</u>
<u ana="#morphemes">...</u>
<spanGrp ana="#glosses">...</spanGrp>
<spanGrp ana="#freeTrans">...</spanGrp>
```

In order to maintain extensibility, TEI-IGT (like Xigt), provides these types as a base structure and allows users to define additional tier types. The values of the `@ana` attribute are references that link to an `<interp>` element somewhere in the corpus. In order to define an new tier type, a new `<interp>` can be added (Listing 6.32).

Listing 6.32: Adding tier types

```
<!-- basic tier types -->
<interp xml:id="phrases">phrases</interp>
<interp xml:id="words">words</interp>
<interp xml:id="morphemes">morphemes</interp>

<!-- new tier type -->
<interp xml:id="syllables">syllables</interp>
```

`<interp>` elements can be grouped into interpretation groups (`<interpGrp>`). Listing 6.33 displays the grouping of `<interp>` elements according to their tier class.

Listing 6.33: Interpretation groups

```
<interpGrp type="segmentationTiers">
        <interp xml:id="text">text</interp>
        <interp xml:id="words">words</interp>
        <interp xml:id="morphemes">morphemes</interp>
</interpGrp>
<interpGrp type="annotationTiers">
        <interp xml:id="glosses">glosses</interp>
        <interp xml:id="freeTrans">free translation</interp>
</interpGrp>
```

Very often, IGT has only one segmentation tier. According to the Leipzig Glossing Rules (Comrie et al., 1982), this tier presents text in the target language, which is segmented down to the morpheme level. While words are separated by whitespace, morphemes are separated by hyphens (or other punctuation for specific kinds of morphemes). Example 6.4 illustrates this structure (taken from the Leipzig Glossing Rules (Comrie et al., 1982, p. 2)).

(6.4)

Gila	*abur-u-n*	*ferma*	*hamišaluǧ*	*güǧüna*	*amuq'-da-č.*
now	they-OBL-GEN	farm	forever	behind	stay-FUT-NEG

'Now their farm will not stay behind forever.'
(Haspelmath, 1993, p. 207)

TEI-IGT, like Xigt, attributes a separate line to each segmentation level. Although this results in redundant information on multiple tiers, this separation is important for four reasons.

1. Only if morphemes and words are separately encoded as segments can they be aligned using ID reference.

2. The form of single morphemes might deviate from how they appear in a word. For example, the nominalized verb *ku* in Abesabesi is *ikòoǹ*, which can be separated into *i-*, *ku*, and *-oǹ*.

3. If different segmentation levels are separated, different alignment types can be chosen (word-alignment vs. morpheme-alignment will be explained later in this section).

4. Toggling on or off different tiers (explained in Section 7.3.3.2) will only be possible if these tiers are defined somewhere.

The separation does not necessarily match the formatting, as the author can still select which levels are shown. The standard three tier layout is thus possible.

The hierarchy of segmentation tiers is expressed through alignment-links. Those are represented as `@corresp` attributes on the lower-hierarchy tiers. The morpheme tier points to the word tier and the word tier to the phrases tier. A custom-tier can be inserted into the core hierarchy, if connected through this linking system.

Listing 6.34: Tier hierarchy

```
<u ana="#phrases" xml:id="ps1">...</u>
<u ana="#words" xml:id="ws1" corresp="#ps1">...</u>
<u ana="#morphemes" xml:id="ms1" corresp="#ws1">...</u>
```

In Xigt, alignment of segments is done through string slicing and joining operations. TEI, on the other hand, relies on nesting and ID reference to align parallel content, which will be used in TEI-IGT. Goodman et al. (2015, p. 465) point out the shortcomings of complex nesting for an IGT format. In response to these shortcomings, the method proposed here only uses flat nesting in each tier to avoid issues with complex nesting but still maintain TEI's logic.

Each tier can be segmented into units that can be referenced from the tier below. Each tier, however, only provides one level of annotation to avoid complex nesting. This kind of segmentation has the same function as the string slicing operations in Xigt. Once again, the arbitrary segment element (`<seg>`) should be used as segmentation unit. As the segments cannot always be attributed to a specific linguistic unit type, the `@type` attribute is not used. A phrases tier can, for example, be segmented into words if the lower level tier contains words. Listing 6.35 displays the phrases and words tiers of Example 6.5 from the Abesabesi Sketch Grammar.

(6.5) *ɛɲɔ* *hu*
snake die
'the snake died' (ibe224-00.029)

Listing 6.35: Segmentation and alignment

```
<u ana="#phrases" xml:id="ps01">
        <seg type="phrase" xml:id="p01">
        <seg xml:id="s01"> ɛŋɔ</seg>
        <seg xml:id="s02"> hu </seg>
    </seg>
</u>
<u ana="#words" xml:id="ws01" corresp="#ps01">
    <seg type="word" xml:id="w01" corresp="#s01"/>
    <seg type="word" xml:id="w02" corresp="#s02"/>
</u>
```

As Listing 6.35 shows, the items in the word tiers use the `@corresp` attribute to reference the segments of the higher level tier. This attribute is meant to establish an alignment between the current element and the higher level item and copy the content of the referenced segment. This is the equivalent of the `segmentation` relation in Xigt. Example 6.6 from the Abesabesi Sketch Grammar also exhibits morpheme segmentation[12]. Listing 6.36 displays how the three segmentation tiers in this example interact. As the word *bàkiès* can be segmented further, the respective segment in the words tier is not simply empty. It contains a segmentation of the word into morphemes (*bà-ki-ès*). The produced morpheme segments can now be referenced in the morphemes tier.

(6.6)

bà-ki-ès	*ɛmũ*
3PL.HUM.SBJ-give-1PL.OBJ	money

‘They gave us money’ (ibe057-00.198)

Listing 6.36: Morpheme alignment

```
<u ana="#phrases" xml:id="ps02">
    <seg type="phrase" xml:id="p02">
        <seg xml:id="s11"> bàkiès </seg>
        <seg xml:id="s12"> ɛmũ </seg>
    </seg>
</u>
<u ana="#words" xml:id="ws02" corresp="#ps02">
    <seg type="word" xml:id="w11" corresp="#s11">
        <seg xml:id="s13"> bà </seg>
        <seg xml:id="s14"> ki </seg>
        <seg xml:id="s15"> ès </seg>
    </seg>
    <seg type="word" xml:id="w12" corresp="#s12"/>
```

[12] In contrast to the conventions of the sketch grammar, the example is given here with the traditional morpheme separation through hyphens.

```
</u>
<u ana="#morphemes" xml:id="ms02" corresp="#ws02">
        <seg type="morpheme" xml:id="m11" corresp="#s13"/>
        <seg type="morpheme" xml:id="m12" corresp="#s14"/>
        <seg type="morpheme" xml:id="m13" corresp="#s15"/>
        <seg type="morpheme" xml:id="m14" corresp="#w12"/>
</u>
```

In Listing 6.36, every alignment also involved copying the content of the aligned segment (or segmenting it further, in the case of *bà-ki-ès*). In IGT, however, aligned segments may have a different form than in the dependent tier. For example, words maybe be referred to in a way that is different from that word in context. Example 6.7 shows an example sentence taken from Childs (2011, p. 204). The verb *wɔ̀ŋɔ́* appears with a high tone in the first tier and is referred to as *wɔ̀ŋɔ̀* with a low tone in the second tier. As Childs dedicates the first tier to the original transcriptions, this high tone might be a tone caused by prosody, a grammatical tone that is not glossed, or even a transcription error.

(6.7) ɲà cè mì kɔ̀ì wɔ̀ŋɔ́
ɲà *cè* *mì* *kɔ̀-ì* *wɔ̀ŋɔ̀*
3PL IPF 1SG PRO-*i* send
'They were sending it (the bread) to me.'

In this case, one word element of the words tier overrides the content of the word in the phrase tier. The two lines would thus look as they do in Listing 6.37. Goodman et al. (2015, p. 472) refer to this process as "shadowing".

Listing 6.37: Shadowing content

```
<u ana="#phrases" xml:id="ps03">
        <seg type="phrase" xml:id="p03">
                <seg xml:id="s21"> ŋà </seg>
                <seg xml:id="s22"> cè </seg>
                <seg xml:id="s23"> mì </seg>
                <seg xml:id="s24"> kɔ̀ì </seg>
                <seg xml:id="s25"> wɔ̀ŋɔ́ </seg>
        </seg>
</u>
<u ana="#words" xml:id="ws03" corresp="#ps03">
        <seg type="word" xml:id="w21" corresp="#s21"/>
        <seg type="word" xml:id="w22" corresp="#s22"/>
        <seg type="word" xml:id="w23" corresp="#s23"/>
        <seg type="word" xml:id="w24" corresp="#s24">
                <seg xml:id="s26">kɔ̀</seg>
                <seg xml:id="s27">ì</seg>
```

```
        </seg>
        <seg type="word" xml:id="w25" corresp="#s25">wɔ̀ŋɔ̀</seg>
</u>
<u ana="#morphemes" xml:id="ms03" corresp="#ws03">
        <seg type="morpheme" xml:id="m21" corresp="#w21"/>
        <seg type="morpheme" xml:id="m22" corresp="#w22"/>
        <seg type="morpheme" xml:id="m23" corresp="#w23"/>
        <seg type="morpheme" xml:id="m24" corresp="#s26"/>
        <seg type="morpheme" xml:id="m25" corresp="#s27"/>
        <seg type="morpheme" xml:id="m26" corresp="#w25"/>
</u>
```

In some cases, segmentation is not possible on the higher level tier. For example, morphemes or words might not be separable as they have merged. The gerund circumfix in Abesabesi, for example, triggers some tonal changes and assimilation so that the resulting gerund cannot be separated through slicing (Example 6.8).

(6.8) *ikòoǹ* *mi* *ɔ́* *ku* *èdɛ́n*
fall\GRND REL 3SG.HUM fall that
'the falling he fell' (ibe229-00.053)

Listing 6.38: Many-to-one alignment

```
<u ana="#phrases" xml:id="ps04">
        <seg type="phrase" xml:id="p04">
                <seg xml:id="s31">ikòoǹ</seg>
                <seg xml:id="s32">mi</seg>
                <seg xml:id="s33">ɔ́</seg>
                <seg xml:id="s34">ku</seg>
                <seg xml:id="s35">èdɛ́n</seg>
        </seg>
</u>
<u ana="#words" xml:id="ws04" corresp="#ps04">
        <seg type="word" xml:id="w31" corresp="#s31"/>
        <seg type="word" xml:id="w32" corresp="#s32"/>
        <seg type="word" xml:id="w33" corresp="#s33"/>
        <seg type="word" xml:id="w34" corresp="#s34"/>
        <seg type="word" xml:id="w35" corresp="#s35"/>
</u>
<u ana="#morphemes" xml:id="ms04" corresp="#ws04">
        <seg type="morpheme" xml:id="m31" corresp="#w31">i</seg>
        <seg type="morpheme" xml:id="m32" corresp="#w31">ku</seg>
        <seg type="morpheme" xml:id="m33" corresp="#w31">oǹ</seg>
        <seg type="morpheme" xml:id="m34" corresp="#w32"/>
```

```
        <seg type="morpheme" xml:id="m35" corresp="#w33"/>
        <seg type="morpheme" xml:id="m36" corresp="#w34"/>
        <seg type="morpheme" xml:id="m37" corresp="#w35"/>
</u>
```

As the example shows, one segment can be aligned with multiple items of the depending tier. Multiple `@corresp` attributes can thus link to the same instance. This type of alignment is referred to as "many-to-one alignment" by Goodman et al. (2015). The simple alignment between two elements of two different tiers is referred to as "one-to-one alignment".

The gerund circumfix is in fact a discontinuous form. TEI has a method for marking different segments as fragments of a single element. The fragments can be connected by using the `@next` and `@prev` attributes to link the following or preceding fragment. Additionally, the attribute `@part` can be used to indicate the position of the fragement: "I" for initial, "M" for medial, and "F" for final. Listing 6.39 displays the first three morphemes in the morphemes tier. In this version, the gerund is marked as discontinuous. The `@subtype` attribute can be used to indicate the morpheme type.

Listing 6.39: Discontinuous morphemes

```
<seg type="morpheme" xml:id="m31" subtype="prefix"
corresp="#w31" part="I" next="#m33">i</seg>
<seg type="morpheme" xml:id="m32" subtype="stem"
corresp="#w31">ku</seg>
<seg type="morpheme" xml:id="m33" subtype="suffix"
corresp="#w31" part="F" prev="#m31">oǹ</seg>
```

Annotation tiers contain items that are aligned with items of segmentation tiers. In contrast to segmentation tiers, they are usually not segmented into smaller units and do not copy the content of aligned segments. In order to create alignments between annotation tiers and segmentation tiers, the `@target` attribute of the `<span>` element can be used. Listing 6.40 displays the encoding of Example 6.5 with all five basic tiers.

Listing 6.40: Glosses and translation

```
<annotationBlock xml:id="igt01">
        <u ana="#phrases" xml:id="ps01">
                <seg type="phrase" xml:id="p01">
                        <seg xml:id="s01">ɛŋɔ</seg>
                        <seg xml:id="s02">hu</seg>
                </seg>
        </u>
        <u ana="#words" xml:id="ws01" corresp="#ps01">
                <seg type="word" xml:id="w01" corresp="#s01"/>
                <seg type="word" xml:id="w02" corresp="#s02"/>
```

```
        </u>
        <u ana="#morphemes" xml:id="ms01" corresp="#ws01">
                <seg type="morpheme" xml:id="m01" corresp="#w01"/>
                <seg type="morpheme" xml:id="m02" corresp="#w02"/>
        </u>
        <spanGrp ana="#glosses" xml:id="gs01" corresp="#ms01">
                <span type="gloss" xml:id="g01" target="#m01">
                        snake
                </span>
                <span type="gloss" xml:id="g02" target="#m02">
                        die
                </span>
        </spanGrp>
        <spanGrp ana="#freeTrans" xml:id="ts01" corresp="#ps01">
                <span type="translation" xml:id="t01" target="#p01">
                        The snake died.
                </span>
        </spanGrp>
</annotationBlock>
```

The example only contains one-to-one alignments. But annotations can also be aligned in a many-to-one or one-to-many relationship. Discontinuous morphemes or words, for example, are both aligned with the same gloss. This one-to-many alignment is shown in Listing 6.41, where all five basic tiers of Example 6.8 are encoded.

Listing 6.41: Annotation of discontinuous and portmanteau segments

```
<annotationBlock xml:id="igt04">
        <u ana="#phrases" xml:id="ps04">
                <seg type="phrase" xml:id="p04">
                        <seg xml:id="s31">ikòoǹ</seg>
                        <seg xml:id="s32">mi</seg>
                        <seg xml:id="s33">ɔ́</seg>
                        <seg xml:id="s34">ku</seg>
                        <seg xml:id="s35">ɛ̀dɛ́n</seg>
                </seg>
        </u>
        <u ana="#words" xml:id="ws04" corresp="#ps04">
                <seg type="word" xml:id="w31" corresp="#s31"/>
                <seg type="word" xml:id="w32" corresp="#s32"/>
                <seg type="word" xml:id="w33" corresp="#s33"/>
                <seg type="word" xml:id="w34" corresp="#s34"/>
                <seg type="word" xml:id="w35" corresp="#s35"/>
        </u>
```

```
        <u ana="#morphemes" xml:id="ms04" corresp="#ws04">
                <seg type="morpheme" xml:id="m31" subtype="prefix"
                corresp="#w31" part="I" next="#m33">i</seg>
                <seg type="morpheme" xml:id="m32" subtype="stem"
                corresp="#w31">ku</seg>
                <seg type="morpheme" xml:id="m33" subtype="suffix"
                corresp="#w31" part="F" prev="#m31">oǹ</seg>
                <seg type="morpheme" xml:id="m34" corresp="#w32"/>
                <seg type="morpheme" xml:id="m35" corresp="#w33"/>
                <seg type="morpheme" xml:id="m36" corresp="#w34"/>
                <seg type="morpheme" xml:id="m37" corresp="#w35"/>
        </u>
        <spanGrp ana="#glosses" xml:id="gs04" corresp="#gs04">
                <span type="gloss" xml:id="g31" target="#m31 #m33">
                        <abbr corresp="glossary.xml#grnd"/>
                </span>
                <span type="gloss" xml:id="g32" target="#m32">fall</span>
                <span type="gloss" xml:id="g33" target="#m34">
                        <abbr corresp="glossary.xml#rel"/>
                </span>
                <span type="gloss" xml:id="g34" target="#m35">
                        <abbr corresp="glossary.xml#thirdSG"/>
                </span>
                <span type="gloss" xml:id="g35" target="#m35">
                        <abbr corresp="glossary.xml#hum"/>
                </span>
                <span type="gloss" xml:id="g36" target="#m36">fall</span>
                <span type="gloss" xml:id="g37" target="#m37">that</span>
        </spanGrp>
        <spanGrp ana="#freeTrans" xml:id="ts04" corresp="#ps04">
                <span type="translation" xml:id="t04" target="#p04">
                        The falling he fell.
                </span>
        </spanGrp>
</annotationBlock>
```

In this example, gloss "g31" aligns to both parts of the gerund circumfix. The `@corresp` can take two values. The example also displays many-to-one alignment. The pronoun *ɔ́* is aligned with two glosses: 3SG and HUM.

Xigt distinguishes three types of links between items of different tiers: `segmentation`, `alignment`, and `content`. The `segmentation` in Xigt is a link that establishes alignment and copies the content of the aligned segment. As TEI-IGT has a separation between

segmentation and annotation tiers, this relation is only used for segmentation tiers. It is represented by the `@corresp` attribute on `<seg>` elements. If a segmentation element is deliberately left empty, this can be indicated by an empty `<gap>` element. The `alignment` relation in Xigt is a link that establishes alignment without copying the content. In TEI-IGT this relation is only used for annotation tiers. It is represented by the `@target` attribute on `<span>` elements. The `content` relation in Xigt is a link between items where no alignment is established but the content is merely copied. Although I do not see any application in the field of grammaticography, it can be expressed through the TEI attribute `@copyOf`.

This section has explained the principles of Xigt and presented its adaptation in TEI: TEI-IGT. The following sections will now expand these basic functions for the purposes of grammaticography. They will explain how TEI-IGT can be used to encode texts and examples.

6.5.5.2 Text database

As mentioned earlier, the text database of a DRG stores transcribed (and annotated) utterances, links them to the recorded utterance and metadata, and groups them into texts. How to store transcribed and annotated utterances has been shown in the previous section. They can be encoded as IGT instances using the proposed TEI-IGT. This section explains how links to recorded utterances can be established and how utterances can be grouped into texts.

TEI proposes a method for storing texts in a corpus. Instead of using the `<TEI>` element as the root of the document, the element `<teiCorpus>` can be used. The TEI corpus can contain multiple `<TEI>` elements. By structuring the text database as a TEI corpus, each text can be encoded as a TEI instance with its own header where metadata can be stored. How to encode metadata in TEI is explained in Chapter 2 of the TEI Guidelines (TEI Consortium, 2020). The encoding of more specific metadata for language corpora is explained in Section 15.2 in the guidelines. For the purpose of grammaticography, the following elements are important to mention: `<particDesc>` for the description of participants, `<settingDesc>` for the description of the setting, `<textDesc>` for the description of how the text has been produced, and `<recordingStmt>` for the description and identification of an associated recording. An example of the structure of the text database with encoded metadata can be seen in the appendix in Listing A.2. The `<recordingStmt>` can also contain a link to the recording of the text. A URL can be stored in the `@source` attribute of the `<media>` element within the `recording` element (Listing 6.42).

Listing 6.42: Linking a recording

```
<teiHeader>
        <fileDesc>
                <titleStmt>
                        <title>The pregnant tortoise</title>
                </titleStmt>
                ...
                <sourceDesc>
                        <recordingStmt>
                                <recording type="audio">
                                        <media url="http://hdl.handle.net/
                                        2196/00-0000-0000-0013-F5CE-D"
                                        mimeType="audio/wav"></media>
                                </recording>
                        </recordingStmt>
                </sourceDesc>
        </fileDesc>
</teiHeader>
```

In order for an application to identify the exact segment of the recording that is attributed to a transcribed utterance, the segment number needs to be indicated inside the `@corresp` attribute of the `<annotationBlock>`. Additionally, the time frame needs to be identified by using the `@start` and `@end` attributes of the `<annotationBlock>`. The utterance can also be attributed to a speaker by using the `@who` attribute. IDs of the speakers defined in the TEI header can be used as values of the @who attribute. Listing 6.43 shows metadata stored on the utterance level.

Listing 6.43: Linking a recording

```
<annotationBlock xml:id="igt01" corresp="texts.xml#ibe119-00.015"
start="16900" end="17980" who="#AA">
...
</annotationBlock>
```

6.5.5.3 Example database

Section 6.3.2.2 has defined an "example" as a transcribed utterance that has been modified and formatted. The example database of a DRG should therefore be able to store examples as IGT instances that link to the original transcribed utterance and also store formatting choices.

The link to the original utterance from the text database can be stored in the `@corresp` attribute of the `<annotationBlock>`. All metadata concerning the recording is not stored in the example database. Instead, it can be retried through this link. Metadata specific

MAIN_TIER	DISLOCATED_TIER
MAIN_TIER	DISLOCATED_TIER
MAIN_TIER	DISLOCATED_TIER
...	...

Table 6.5: Example Grid

to the example, however, is discussed in the last paragraph of this section.

A peculiarity of the example database is the storage of formatting choices. Formatting choices comprise all meaningful decisions made about the layout, the highlighting, and other enrichments of the example. Formatting choices that do not convey a meaning are not considered.

Layout IGT layout involves the number of tiers displayed, their alignment type, order, and position. The basic structure of an IGT instance consists of several different tiers, one below the other, with their items aligned horizontally. In many examples, however, additional tiers can be found to the right of one of the stacked tiers. Frequently dislocated tiers include phonetic transcriptions or notes and also translations (Example 6.9)

(6.9) (a) *hɔ̃ ebid → [hẽbid]*
harvest palm_fruit
'harvest palm fruits'
(Abesabesi: ibe350-00.042)

(b) *So bwarjga='a jeina* ***aara-vealar*** *Muusaa*
1s eye+ J.see.CV out-V.go.WP Musa
Musa saw me and left. [core chained]
(Nichols, 2011, p. 681)

(c) Liguiyta yachachimanga.
ligi-y-ta yača-či-ma-nqa 'He will teach me to read.'
read-INF-OBJ learn-CAUS-→1-3FUT
(Weber, 2005a, p. 453)

In theory, these dislocated tiers can be any tier defined in TEI-IGT, although it may not be aligned with segments. In order to allow the author the freedom to dislocate tiers but not permit an overly complicated structure, the example layout will be confined to a grid structure of n * 2 tiers (Table 6.5). The identifier is part of the exemplar, but the reference to the original utterance is part of the example. Its layout, however, can be unified as it is not meaningful.

An example in the example database can contain more tiers than are actually displayed in the GET. The author needs to select the tiers that should be displayed and their order in the grid. There should be a default format such as the frequent *morphemes-glosses-translation* triade but, for each example, the author needs to be able to chose a different

layout. In TEI-IGT, the `@rend` attribute is used to store formatting choices. The attribute's values are key-value expressions in the form `key(value)`. There are three available keys: `layout`, `align`, and `style`. While `layout` and `align` can only be added to the tier level (`<u>` and `<spanGrp>`), `style` can also be added on the item level (`<seg>` and `<span>`). To select a tier for display, the tier needs to be assigned a position in the grid. Therefore, the key `layout` is used together with two integer parameters – the first one assigning the row and the second one the column. `layout(2,1)` is, for example, the second main tier and `layout(4,2)`, the fourth dislocated tier. The standard *morphemes-glosses-translation* triade would, for example, be encoded as in Listing 6.44. All tiers that have not been chosen for display could still be toggled out by the reader.

Listing 6.44: Morphemes-glosses-translations layout

```
<annotationBlock xml:id="igt05" corresp="texts.xml#ibe119-00.015"
start="16900" end="17980" who="#AA">
        <u ana="#phrases" xml:id="ps05">
                ...
        </u>
        <u ana="#words" xml:id="ws05" corresp="#ps05">
                ...
        </u>
        <u ana="#morphemes" xml:id="ms05" corresp="#ws05" rend="layout(1,1)">
                ...
        </u>
        <spanGrp ana="#glosses" xml:id="gs05" corresp="#ms05" rend="layout(2,1)">
                ...
        </spanGrp>
        <spanGrp ana="#freeTrans" xml:id="ts05" corresp="#ps05" rend="layout(3,1)">
                ...
        </spanGrp>
</annotationBlock>
```

Table alignment IGT is in fact a specialized version of a table and is rendered as such. In order to display the alignment of segments and annotations, aligned items are displayed in the same column. The Leipzig Glossing Rules (Comrie et al., 1982) suggest the word-alignment type where each word is displayed in a separate column. Smaller segmentation (morphemes) are separated within the column through hyphens. As glosses are usually aligned to morphemes, they also are separated by hyphens (Example 6.10).

(6.10)

i-ku-oǹ	*mi*	*ɔ́*	*ku*	*èdɛ́n*
GRND-fall-GRND	REL	3SG.HUM	fall	that

'the falling he fell' (ibe229-00.053)

As shown earlier, words might look different than the sum of their morphemes, which is why this alignment type is not always preferable. In this case, the word appearance is not shown and native speakers might not even recognize the word in question. The Leipzig Glossing Rules propose a solution for this case (Example 6.11). A backslash (\) is used to gloss inseparable morphemes. The segmentation of morphemes, however, is not shown.

(6.11)

ikòoǹ	*mi*	*ɔ́*	*ku*	*èdɛ́n*
fall\GRND	REL	3SG.HUM	fall	that

'the falling he fell' (ibe229-00.053)

An alternative is the morpheme-alignment type. This type is, for example, used by FLEx when displaying IGT. Each word gets a column and each morpheme gets a subcolumn underneath the word alignment. Example 6.12 illustrates this alignment type.

(6.12)

ikòoǹ			*mi*	*ɔ́*	*ku*	*èdɛ́n*
i-	ku	-oǹ	mi	ɔ́	ku	èdɛ́n
GRND	fall	GRND	REL	3SG.HUM	fall	that

'the falling he fell' (ibe229-00.053)

The decision whether to follow the Leipzig Glossing Rules standard or to use morpheme alignment should be up to the author. In TEI-IGT, the formatting expression `align(anchor)` can be attributed to one of the segmentation tiers in order to select which tier contains the smallest alignment unit. If this expression is used on the word tier, word-alignment is selected, as proposed by the Leipzig Glossing Rules (as in Example 6.10). If used on the morpheme tier, the morphemes are the smallest alignment unit. Thus, morphemes will be printed in separate columns and words, in super columns (as in Example 6.12). Listing 6.45 shows the encoding of an example with morpheme alignment. Theoretically, the smallest alignment unit could also be more complex units, such as segments on the phrase level.

Listing 6.45: Encoding of morpheme alignment

```
<annotationBlock xml:id="igt05" corresp="texts.xml#ibe119-00.015"
start="16900" end="17980" who="#AA">
        <u ana="#phrases" xml:id="ps05">
                ...
        </u>
        <u ana="#words" xml:id="ws05" corresp="#ps05" rend="layout(1,1)">...
        </u>
        <u ana="#morphemes" xml:id="ms05" corresp="#ws05"
        rend="layout(2,1) align(anchor)">...
        </u>
        <spanGrp ana="#glosses" xml:id="gs05" corresp="#ms05"
        rend="layout(3,1)">...
```

```
        </spanGrp>
        <spanGrp ana="#freeTrans" xml:id="ts05" corresp="#ps05"
        rend="layout(4,1)">...
        </spanGrp>
</annotationBlock>
```

Highlighting Authors use different ways of letter styling to guide the reader's attention. Letter styling can come in the form of italics, underlined or bold elements, all capital letters, small capital letters, fonts or color. As the format is less concerned with the form of elements, it will not be of concern how certain elements are styled. In the same way, the styling of an entire tier with the intention of differentiating tiers is not important to this format as it does not convey meaning. Styling elements according to their type is also not of concern for similar reasons. For example, gloss abbreviations may all be rendered in small capitals. This can depend on the application or it can be defined in a stylesheet. Styling single elements regardless of their type, however, can convey meaning if it indicates the point of discussion. This function will be referred to as "highlighting". Example 6.13 displays an exemplar taken from the Ingush Grammar (Nichols, 2011, p. 269). In order to identify the discussed construction (simultaneous converb (CVsim) + verb in present tense (PRS)), two words are printed in boldface. This becomes even more important in more complex utterances.

(6.13) *Gour* ***kerchazh*** ***ull***
horse roll.CVsim lie.PRS
'The horse is rolling around on its back.' (Nichols, 2011, p. 269)

In order to highlight a segment or annotation, the format expression `style(hi)` is used as a value for the `@rend` attribute. Listing 6.46 displays the encoding of Example 6.13.

Listing 6.46: Encoding of highlighting

```
<u ana="#words" xml:id="ws06" corresp="#ps06" rend="layout(1,1) align(anchor)">
        <seg type="word" xml:id="w51">
                Gour
        </seg>
        <seg type="word" xml:id="w52" rend="style(hi)">
                kerchazh
        </seg>
        <seg type="word" xml:id="w53" rend="style(hi)">
                ull
        </seg>
</u>
```

As can be seen in the Ingush example, the highlighting can also have the secondary function of indicating constituency. In this case, it could also be encoded as a `<span>` in a separate annotation tier. This annotation tier is not rendered within the grid but in the form of highlighting (Listing 6.47)

Listing 6.47: Encoding constituency through highlighting

```
<u ana="#words" xml:id="ws06" corresp="#ps06" rend="layout(1,1) align(anchor)">
        <seg type="word" xml:id="w51">
                Gour
        </seg>
        <seg type="word" xml:id="w52">
                kerchazh
        </seg>
        <seg type="word" xml:id="w53">
                ull
        </seg>
</u>
<spanGrp ana="#constructions" xml:id="cs06" corresp="#ws06">
        <span type="construction" from="#w52" to="#w53" rend="style(hi)"/>
</spanGrp>
```

A more common way to mark constituency is to surround constituents with brackets. These brackets can have labels to indicate the phrase type. This can also be encoded through elements and the format expression `style(brackets)` (Listing 6.48). The label is the content of the `<span>` element.

Listing 6.48: Encoding constituency through brackets

```
<spanGrp ana="#constructions" xml:id="cs06" corresp="#ws06">
        <span type="construction" from="#w52" to="#w53" rend="style(brackets)">
                VP
        </span>
</spanGrp>
```

Other enrichments Three other common kinds of example enrichment worth encoding are grammaticality judgments, co-reference subscripts, and notes on items. Grammaticality judgments indicate the degree of grammaticality speakers attribute to an utterance. Utterances judged ungrammatical might, for example, be used as negative evidence in a grammatical description. In practice, only ungrammatical utterances and those that are judged to be uncommon are marked using an asterisk or a question mark. These can be encoded by using the format expressions `style(ungramm)` (ungrammatical) and `style(uncom)` (uncommon).

Subscript numbers or letters can be used on segments to indicate co-referentiality. If, for example, two pronouns refer to the same entity, both pronouns get the same sub-

scripted letter/number. Referents can also be encoded as annotation tiers. An annotation tier that is to be rendered as subscript should receive the `@rend` attribute `style(subscript)`. A span can refer to one referent and have multiple targets (`@target`). The content of the span will be rendered as subscript. The following examples of Abesabesi logophorics use subscript letters to indicate co-referentiality (Example 6.14 a) and lack of co-referentiality (Example 6.14 b).

(6.14) (a)

ɔ́	*mò*	*àu*	*á*	*kɔ*	*lɛ́tà*
3SG.HUM	say	3SGHUM.LOG	PROG	write	letter

'S/he$_i$ said s/he$_i$ is writing a letter' (ibe203-00.090)

(b)

ɔ́	*mò*	*mí*	*ɔ́*	*á*	*kɔ*	*lɛ́tà*
3SG.HUM	say	COMP	3SGHUM	PROG	write	letter

'S/he$_i$ said s/he$_j$ is writing a letter' (ibe203-00.096)

Listing 6.49 displays the encoding of referents in both examples. Co-referentiality is encoded by one span with two targets and lack of co-referentiality is encoded by two spans with one target each.

Listing 6.49: Encoding referents

```
<!-- Co-referentiality -->
<spanGrp ana="#referents" xml:id="rs07" corresp="#ws07" rend="style(subscript)">
        <span type="referent" target="#w61 #w63">i</span>
</spanGrp>
<!-- Lack of co-referentiality -->
<spanGrp ana="#referents" xml:id="rs08" corresp="#ws08" rend="style(subscript)">
        <span type="referent" target="#w61">i</span>
        <span type="referent" target="#w64">j</span>
</spanGrp>
```

Explanatory notes on segments or annotations are equal to (foot)notes in the prose or other parts of the grammar. They will therefore be likewise encoded as a `<note>` element within the `<seg>` or `<span>`. Notes are discussed in Section 3.8 of the TEI guidelines (TEI Consortium, 2020).

Metadata As mentioned at the end of Section 6.5.5.2, most of the metadata is stored in the text database on the text or utterance level. In the example database, only one kind of metadata needs to be stored: keywords. Examples can be enriched by attributing keywords to them for indexing. This way, examples that contain a specific phenomenon can be searched. TEI does not permit the `<index>` element in an annotation block. Therefore, the index will be added to the first utterance element. Within the index, multiple `<term>` elements can be used as keywords (Listing 6.50)

Listing 6.50: Index in examples

```
<annotationBlock xml:id="igt01" corresp="texts.xml#ibe119-00.015"
start="16900" end="17980" who="#AA">
        <u ana="#phrases" xml:id="ps01">
                <index>
                        <term ref="terminology.xml#focus"/>
                        <term ref="terminology.xml#independentPronoun"/>
                </index>
                <seg>...</seg>
                ...
        </u>
        ...
</annotationBlock>
```

6.5.6 TEI adjustments

TEI-Grammar has been designed to make use of only existing TEI elements. In some cases, alternatives have been proposed, as the use of existing elements is not optimal and could rather be regarded as a work-around. This section therefore contains some proposals for new elements or attributes that could improve the encoding of grammars with TEI.

TEI-Grammar uses the text division element (`<div>`) to represent the central unit of grammar writing – grammar entries. Although grammar entries are a kind of text division and similar to chapters and sections, a more specialized tag could guarantee a more rigid structure and thus control the encoding better. A separate tag such as `<gramEntry>` could be used to allow only certain elements inside a grammar entry. A similar case is the `<entry>` element for lexical resources. This element also represents a kind of text division, but it is only allowed to contain a limited amount of element types and therefore sets limits to how a dictionary entry is structured.

As mentioned in Section 6.5.3.3, segmental units of the model.segLike class like `<seg>`, `<w>`, or `<cl>` are only designed to contain the form part of a segment. Using the element `<entry>` instead is not an alternative, as in grammatical descriptions, these segmental units are not introduced as a new entry but instances of the entry used in a text. Segmental units rather refer to an entry. The goal is not to establish a new entry with as much data about it as possible, but merely to indicate which lexeme, word, morpheme, etc. the discussion is about. Therefore, segmental units should be allowed to contain a form and meaning side like `<entry>` does. An example for this has been given at the end of Section 6.5.3.3.

The use of segmental units other than `<seg>` has been rejected in TEI-Grammar because not enough specialized segment categories are available for a grammatical description. In order to use the specialized elements such as `<m>` or `<phr>`, TEI would need to provide elements for more units such as a construction, an intonation contour, or a

phoneme.

Another method to represent segment categories would be to continue using the arbitrary segment (<seg>) but to connect the segment type to the used terminology. In the same way, parts of speech or other grammatical labels in the dictionary could link to terms in the terminology. This way, the analysis presented in the grammatical description could be synchronized with the structure of the lexical database. More coherence in terms of analyses has already been postulated in Maxim 5. For this endeavor, the elements <seg> and <gram> would need to have an attribute like @typeRef that contains a reference to a term that labels the segment or category type. Likewise, @subtypeRef could contain a reference to subtype labels, such as "prefix" of "suffix". Listing 6.51 displays these attributes on a <seg> element and Listing 6.52 displays them on <gram> elements in a lexical entry.

Listing 6.51: @typeRef and @subtypeRef attributes on <seg> element

```
<seg typeRef="terminology.xml#morpheme"
subtypeRef="terminology.xml#circumfix">i- -oǹ</seg>
```

Listing 6.52: @typeRef attribute on <gram> elements

```
<entry xml:lang="ibe" xml:id="odʒíbèrè_003a55e8-367f-4910-a2e7-e66b8adbdf76">
        <form type="lemma">
                <pron notation="ipa">odʒíbèrè</pron>
        </form>
        <sense>
                <gloss>gorilla</gloss>
                <cit>
                <quote xml:lang="en">
                        gorilla
                </quote>
                <gramGrp>
                        <gram typeRef="terminology.xml#pos">Noun</gram>
                        <gram typeRef="terminology.xml#suffixVowelHarmony">
                                <foreign xml:lang="ibe">na</foreign>
                        </gram>
                </gramGrp>
                </cit>
        </sense>
</entry>
```

Throughout TEI-Grammar, the attribute @corresp is used very often and in diverging contexts. In most cases, it is used to establish links to entries in a database. For example, it is used by the element <abbr> to reference an abbreviated term from a database and it is used by linguistic units to reference lexical entries. Within TEI-IGT, it is also used for

alignment. A big problem is that segmental units in interlinear examples use the attribute for alignment to other segmental units and for references to lexical entries. Multiple URIs can be stored in the attribute and an application could distinguish the two types of links by the structure of the URI, but this is certainly not the most elegant method. `@corresp` has a very general use in TEI and is used for different kinds of "correspondences" and, for many elements, it is the only appropriate possibility to store links to other elements. An alternative would be to allow for the use of `@ref` for the segmental units. In fact, segmental units in grammars contain a similar kind of reference to their lexical entries as terms do to terminological entries in an ontology.

Chapter 7

Developing a Grammar Exploration Tool

The data format proposed in the previous chapter is not more than an empty husk as long as it is not tested by filling it with data and by letting applications interact with it. The Abesabesi Grammar Sketch was used to test the encoding of a grammar sketch (Section 6.4). This chapter will now test how encoded data can be read by an application. As the development of a grammar writing tool (GWT) is a project beyond the scope of this PhD, only a GET was developed to test the interaction between the proposed format and an application. In order to differentiate the exemplary GET developed within this project from the concept of GETs, it will be referred to as the "Abesabesi GET". The Abesabesi GET can be accessed at `https://abesabesi.cceh.uni-koeln.de/`. What is meant by a GET was already clarified in Section 4.5 and its structure was sketched out in Section 5.4.2. This sketch will be used and extended to present an outline of a GET's structure, on which the development of the application is based (Section 7.1). After Section 7.2 presents the technologies used for the development, Section 7.3 documents the implementation of the proposed features. Section 7.4 presents features that were not implemented. And finally, Section 7.5 discusses the shortcomings of the format that could be detected while developing the application and proposes adjustments.

7.1 An outline for the GET

Since a rough structure of a GET has already been sketched out in Section 5.4.2, this Section will only elaborate on a GET's structure by considering more of the features proposed in Chapter 5 and produce an outline that was used as a blueprint for the development of the Abesabesi GET. In order to give a first impression of the Abesabesi GET's extent, the following outline lists will indicate which features have been implemented (I), which ones have been implemented as a mockup without functionality (M), and which ones have not been implemented (N)

A GET is aimed at reading the grammar documents and displaying the encoded gram-

matical description in a reader-friendly way. The Abesabesi GET displays only one grammatical description: the Abesabesi Sketch Grammar. It therefore follows the second approach proposed in Section 5.4.2 – a GET specifically designed for one grammatical description. Since the Abesabesi GET was developed as a web application, the following outline is structured accordingly. Like in Section 5.4.2, it is separated into the parts "general website functionalities", "grammar entries", "access to the databases", and "navigation tools". The same division will be used later in Section 7.3, where the implementation of the features will be discussed.

- General website functionalities
  - Structure the website adequately
    - A homepage with quick background information about the language and the grammar (I)
    - A place where the grammatical description can be read (I)
    - A place or places where the different databases can be accessed (I)
    - A place where detailed background information is given about the website and its content ("about pages") (I)
  - Provide additional functionalities
    - A registration and login facility where users could register and log in if access is restricted (M)
    - Navigation tools that help the reader find information within the grammatical description or the databases (I)
    - Offline usability to enable reading without an internet connection (N)

The general website functionalities for the GET have been taken entirely from Section 5.4.2. They include a suggestion for the structure of the website and the provision of tools that should be accessible from any part of the website.

- Grammar entries
  - Read grammar document (I)
  - Display grammar entries as separate hypertext pages (I)
  - Provide ways to access the preceding, following, upper, and lower pages. (I)
  - Display and distinguish the elements of the micro- and nanostructure. (I)
    - Format tables, lists, figures, and interlinear examples (I)
    - Display interlinear examples in a default way and allow toggling on/off additional tiers (N)
    - Display metadata tags, such as the keywords used as indexes (I)
    - Use a specific format for all elements of the nanostructure that indicate they can be clicked – except for linguistic units (I)

    - Format linguistic units in italics and translations/glosses in quotation marks (I)
    - Provide tooltips for abbreviations to display their expansion (I)
  - Resolve references by establishing hyperlinks
    - Links to other elements in the grammatical description (I)
    - Links between linguistic units and their dictionary entries (I)
    - Links from examples to access metadata and recordings (M)
    - Links from examples to texts (I)
    - Links from terms and abbreviations to ontologies (I)
    - Links to bibliography entries (I)
  - Provide a tab for each grammar entry to see its edits and the associated authors. (M)

Grammar entries need to be extracted from the main grammar document and displayed as separate hypertext pages. Internal navigation tools could allow navigation between grammar entries. All content of the entries (micro- and nanostructure) need to be formatted in appropriate ways to distinguish the different elements. How exactly these elements are displayed will be discussed in Section 7.3.2. References encoded in the grammar documents need to be resolved and represented as hyperlinks to grant easy access to databases or within the grammatical description. Additionally, tabs could be provided for each grammar entry to see its version history: when and who edited which part of the entry?

- Access to the databases
  - Text database
    - Display texts (with glosses and translation) (I)
    - Provide a way to access metadata (M)
    - Provide a way to access the respective recording (M)
  - Example database
    - Display examples from the example database (I)
    - Provide a way to access metadata (M)
    - Provide a way to access the respective recording (M)
    - Display the associated keywords (I)
    - Provide a link to the associated grammar entries (I)
    - Provide a link to the associated text if applicable (I)
  - Lexical database
    - Display all lexical entries (I)
    - Provide a way to search for a specific entry (I)

        * Provide links to all grammar entries where a lexical entry appears. (I)
    - Terminology database
        * Display all terms of the database (I)
        * Simulate the hierarchical structure of each ontology (I)
        * Provide a way to access the term's definition and the associated grammar entries/examples (I)
    - Abbreviation database
        * Display all abbreviations and their expansions (I)
    - Bibliography database
        * Display all bibliographic entries (I)
        * Provide an output of the entries in several bibliographic formats (M)

Access to the databases is important for all readers that prefer a database-driven approach to reading a grammar. Each database could be rendered as one hypertext page. These database pages and their implementation in the Abesabesi GET will be discussed in Section 7.3. Databases can also act as navigation tools that help find specific grammar entries. For instance, all grammar entries could be searched that discuss a specific lexical entry. Accessing the lexical database can help find the entry and lead to the grammar entries containing the lexical entry. Therefore, it is important to provide plenty of connections between the databases and the grammatical description[1]. As discussed in Section 6.5, various kinds of references can be encoded in the the grammar documents. These references might lead from the main grammar document to the databases or from one database to another. A GET could utilize these encoded references and display them as hyperlinks that guide the reader quickly to the referenced database entries. Database entries, in turn, could also refer back to all grammar entries where they are used.

- Navigation tools
    - TOC
        * Display the headings in a nested, numbered list for the traditional approach (I)
        * Provide a way to fold/unfold deeper levels (I)
        * Represent relations between entries in a form/function approach (N)
        * Display domains and their headings ordered according to their priority (N) (dynamic approach)
        * Format read entries differently from unread entries (N)
    - Index
        * Display all indexes in a list (I)

[1] Possible links between grammar documents were displayed in Figure 5.1

    - Link the used indexes to their respective terms in the terminology database (I)
    - Provide a way to access associated grammar entries or examples (I)
  - Search functionalities
    - Provide a full text search form (I)
    - Provide advanced search forms that allow for contextualized and combined queries as well as for regular expressions. (M)
  - Version history
    - Provide a way to access the current as well as previous versions (M)
    - Provide a reminder of which version is currently accessed (M)
  - Layer selection
    - Provide a way to select a specific layer of the grammar. (N)

Navigation tools help the reader find specific information in the right grammar entry or in one of the databases. How navigation tools work in paper grammars has been discussed in Section 6.3.1.1. GETs could provide a TOC, an index, search functionalities, a version history, and a layer selection as navigation tools. A TOC could be designed in different ways. The different design approaches are related to different concepts of a grammatical description's linearity (Section 5.3.7). Theoretically, the different approaches could all be implemented and the readers could chose a preferred design. However, this would require the encoding of all metadata necessary to each approach (headings, domains, priority, etc.). An index should list all indexes and link to the grammar entries or examples that are tagged with the respective index term. They could also be equipped with a link to the definition of the respective term. Search functionalities are unique to digital grammars and include full text search and advanced search for contextualized and combined queries. Version history is important to ensure the persistence of the grammar and to attribute different edits to different authors in case of distributed authorship (Section 5.3.4). Layer selection could help the reader to select a layer if the grammatical description is encoded in layers. These layers could be different varieties of a language or different metalanguages (Section 5.2.8). All of these navigation tools and their implementation are discussed in Section 7.3.4.

7.2 Development of the Abesabesi GET

As outlined in the previous section, the Abesabesi GET has been planned as a web application that can read the grammar format and display the grammatical description within a web browser. For the development of the GET, I used the web application platform eXist-db (EXIST-DB, 2012), which has several advantages for this project.

eXist-db is a NoSQL native XML database that has the document as its atomic unit. It has primarily been designed to handle data-oriented and text-oriented XML documents

while offering a platform to easily develop web applications. It is an open source software that has entirely been built around XML technologies. Thus, it has been used for a range of academic projects, especially in the field of digital humanities, where XML and especially TEI documents are popular data formats. Instead of a proprietary database query language, eXist uses XQuery. This standardized query language has been developed by the W3C (Section 6.2.1) (W3C, 2017). Code written in XQuery can therefore not only be used in eXist-db but in any other platform that supports XQuery. Although referred to as a "query language", XQuery is a complete functional programming language that can be used to program entire applications (Siegel & Retter, 2015, p. 2).

The eXist database stores the TEI grammar documents and other files, such as pictures, audio, and video recordings. HTML pages form the static base of the web-pages that are enhanced by XQuery functions. All XQuery functions concerning the interactions between the HTML pages and the TEI documents are defined in six XQuery files. Each XQuery document represents a separate module and accesses different grammar documents. All functions related to the grammar entries and the TOC are defined in `grammar-entry.xql`. It accesses the main grammar document. `dict.xql` accesses the lexical database and contains functions for the dictionary page. All IGT-related functions are defined in and `igt.xql`, which accesses the example and text databases. `bib.xql` accesses the bibliography database and contains all functions for the bibliography access page. Terms and abbreviations are managed by `terms.xql`, which accesses the terminology and abbreviation databases. Finally, `search.xql` handles all search requests and also accesses the main grammar document. A function is usually called within an empty HTML element through the HTML templating framework. It then queries the TEI files and returns XHTML snippets that contain the desired data from the grammatical description. Specific entries of the grammatical description, the lexical database, or the terminology database are accessed through a parameter in the URL. A URL requesting a specific grammatical entry could for example look like this:

`https://abesabesi.cceh.uni-koeln.de/grammar-entry.html?section=ch2`.

The parameter "section" has the value "ch2", which is the identifier of a grammar entry (`<div>` element) within the TEI document. This URL first accesses the `grammar-entry.html`, which in turn contains a call to a function processing the specific grammar entry. The function then accesses the parameter of the URL in order to retrieve data of the requested grammar entry.

Figure 7.1 displays the application logic and the interaction between the HTML files, the XQuery files, and the TEI files. The following section is concerned about the implementation of additional features. These features are mainly defined through XQuery functions within the six XQuery files.

The developed GET application runs on a virtual machine provided by the Regionales Rechenzentrum der Universität Köln (`https://rrzk.uni-koeln.de/`). This virtual machine is used by the Cologne Center for eHumanities (`https://cceh.uni-koeln.de/`) and made available for the grammar application. This way, it receives some basic main-

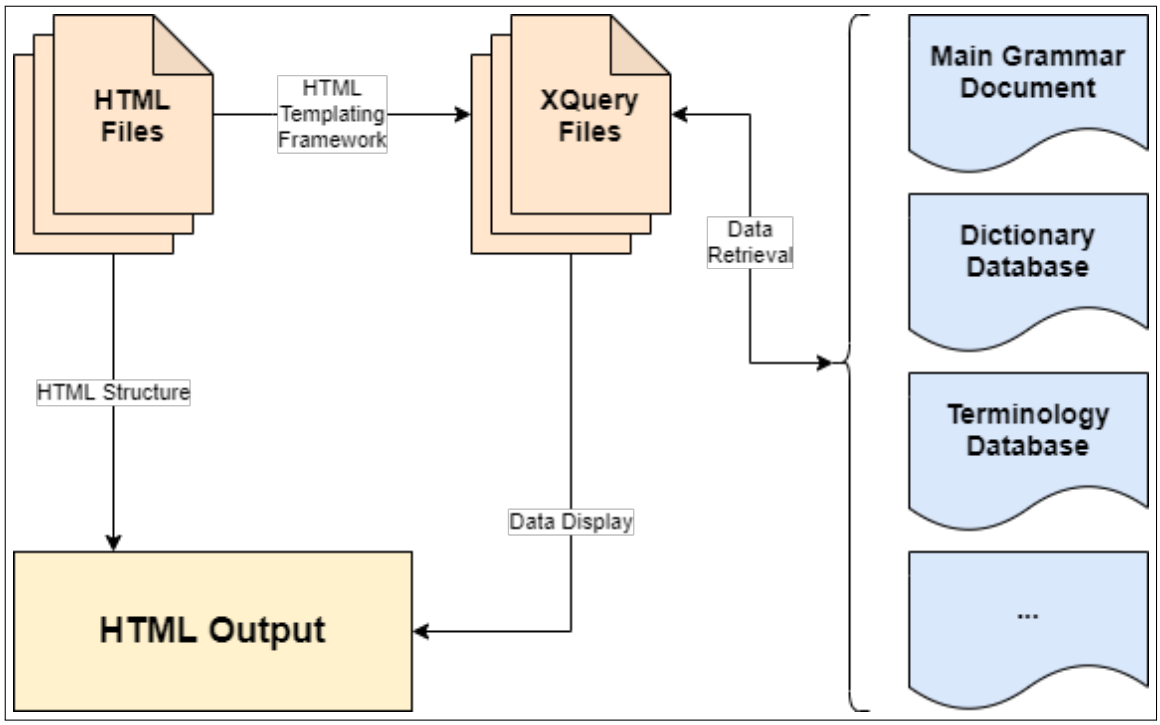


Figure 7.1: Abesabesi GET application logic

tenance to allow readers to examine the application and for people of the language community to use the dictionary. Yet, I am aware of the fact that its longevity cannot be entirely secured and the application might not be available in a few years.

7.3 Implementing the features

Before implementing all features outlined in Section 7.1, it is important to reconsider the reason for developing the GET. The following list presents the two main reasons to develop the GET and what exactly will be important for the evaluation of the GET and the format.

1. Exemplify a GET and its functionalities
   (a) Test whether the functionalities can be implemented
   (b) Test whether the functionalities improve the reading experience
2. Test the proposed grammar format
   (a) Test the interaction between format and application
   (b) Test whether all content and the represented relations can be extracted from the format
   (c) Find out possible shortcomings of the format

The first reason is to exemplify a GET and its functionalities. This is important as the feasibility of the development with features proposed in the previous chapters needs to be proven. Features that are not feasible are only conceptual until someone manages to implement them. Needless to say, some features will not be implemented within this GET for reasons of time, but their feasibility within a larger scope will be discussed in Section 7.4. Furthermore, the development of this GET sets an example of how to develop such an app and could be inspiration for a new development project. It may also serve as a basis for discussion in further theoretical discourse about digital grammaticography. Testing the reading experience (Reason 1b) within the scope of this project is very subjective as it has been done only by myself. This is also due to time limits. Valuable quantitative testing remains an idea for future research.

As the proposal for a grammar format for DRGs has been chosen to be the focus of this thesis, the main reason to develop a GET is to test the proposed grammar format (Reason 2). Besides storing the grammatical description, the grammar format also needs to be machine-readable and applications need to be able to interact with the format (Reason 2a). A GET primarily extracts data and displays it. The grammar format is thus only a suitable format if a GET can extract all stored information and the represented structure of the grammatical description (Reason 2b). Moreover, a GET formats the encoded grammatical description and displays it in the reader-friendly way. If a desired layout is not possible because different structural elements are not distinguished by the format, this is a shortcoming of the format. The development of the GET thus facilitates the detection of the format's shortcomings. Shortcomings and how they can be adjusted will be discussed in Section 7.5.

These reasons for the development of a GET also indicate what will be important for the evaluation of the application and the format. This section now discusses how the proposed features have been implemented and evaluates them one by one. The following Section 7.4 will discuss the features that could not be implemented and their feasibility in a project of a larger scope. Some features have only been implemented as a "mockup". They look like a functioning feature but do not have the functionalities when clicked or hovered upon. They were indicated by an (M) in the outline presented in Section 7.1. Section 7.5 then discusses the shortcomings of the format that have been detected from the development of the GET and how they can be adjusted.

7.3.1 General website functionalities

As the main focus is the evaluation of the format, general website functionalities will not be discussed in detail or evaluated. Only a short overview about the web site's general structure will be given. Login functionalities and offline usability have not been implemented as they do not allow conclusions concerning the format.

The website has three main layouts. The homepage layout (Figure 7.2), the grammar layout (Figure 7.3), and the dictionary layout (Figure 7.5). All layouts are mainly conceptualized for a computer screen, as I assume that academic work is mostly done on

the computer. However, a grammar should be accessible to as many people as possible (value accessibility in Section 5.2.2) and, therefore, the GET should also be accessible from smaller devices, such as a mobile phone. This GET has a responsive design so that all layouts have an equivalent for smaller devices. Yet, the equivalents for smaller devices have not been elaborated on as much as the layouts shown here.

The homepage layout is used for the homepage, the about and help pages, and the navigation tools. It includes a navigation bar to the left, a big picture of the town Ikaram, basic facts about Abesabesi to the right, and a more conspicuous navigation bar below the picture. The content of the page is below this navigation bar. The navigation bar on the left side is included in all layouts to grant access to all parts of the website from every page. It contains a mockup of a login feature on the top. In this scenario, the user is logged in, as the profile picture is shown and the user is welcomed. This page layout is used for pages which contain short information and welcome the reader. The Abesabesi GET contains five about pages: “About the language”, “About Ikaram”, “About the grammar” (which functions as a help page), “About the documentation project”, and “Acknowledgments”. As the “About the language” page overlaps with the content of the grammar’s introductory chapters, it contains a short summary and links to the introductory chapters related to Abesabesi, such as “Genetic affiliation”, “Sociolinguistics”, and “Surrounding languages”.

The grammar layout is less colorful and is used for all pages containing a grammar entry. The focus here is to accommodate a larger amount of content and not to distract the reader from the main content. A toolbar is situated on the right side. It includes the tags for quality assessment and relative importance on the top, the keywords used as indexes below and access to a few tools at the bottom. In order to facilitate an undisturbed reading-flow with more space for the main content, this toolbar can be hidden by clicking the “Hide toolbar” button (see Figure 7.4). All other peculiarities of this layout will be discussed in Section 7.3.2.

The dictionary layout is used for all pages containing a lexical entry. This layout will be explained along with the dictionary page in Section 7.3.3.3.

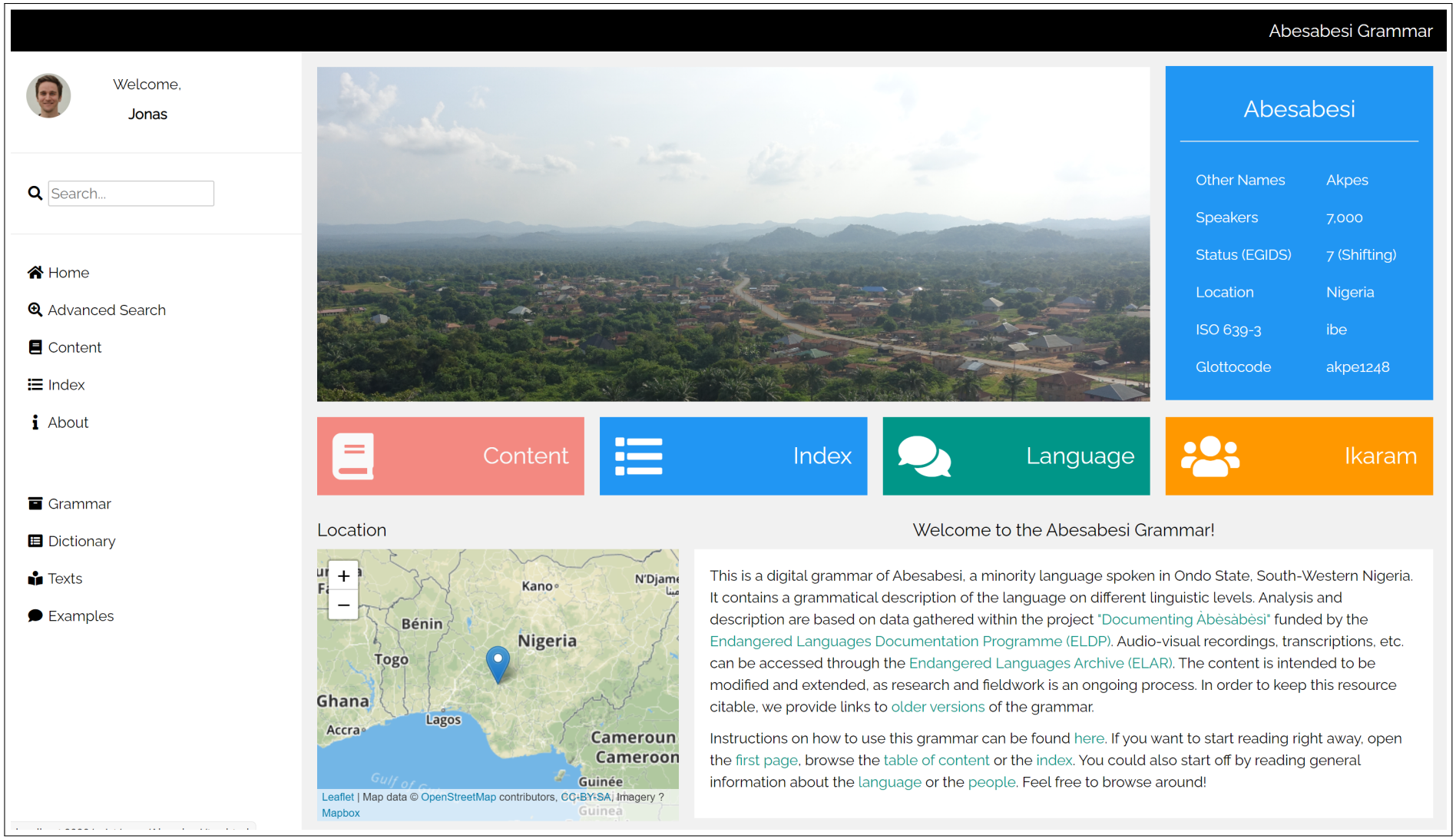


Figure 7.2: Homepage of the Abesabesi Grammar

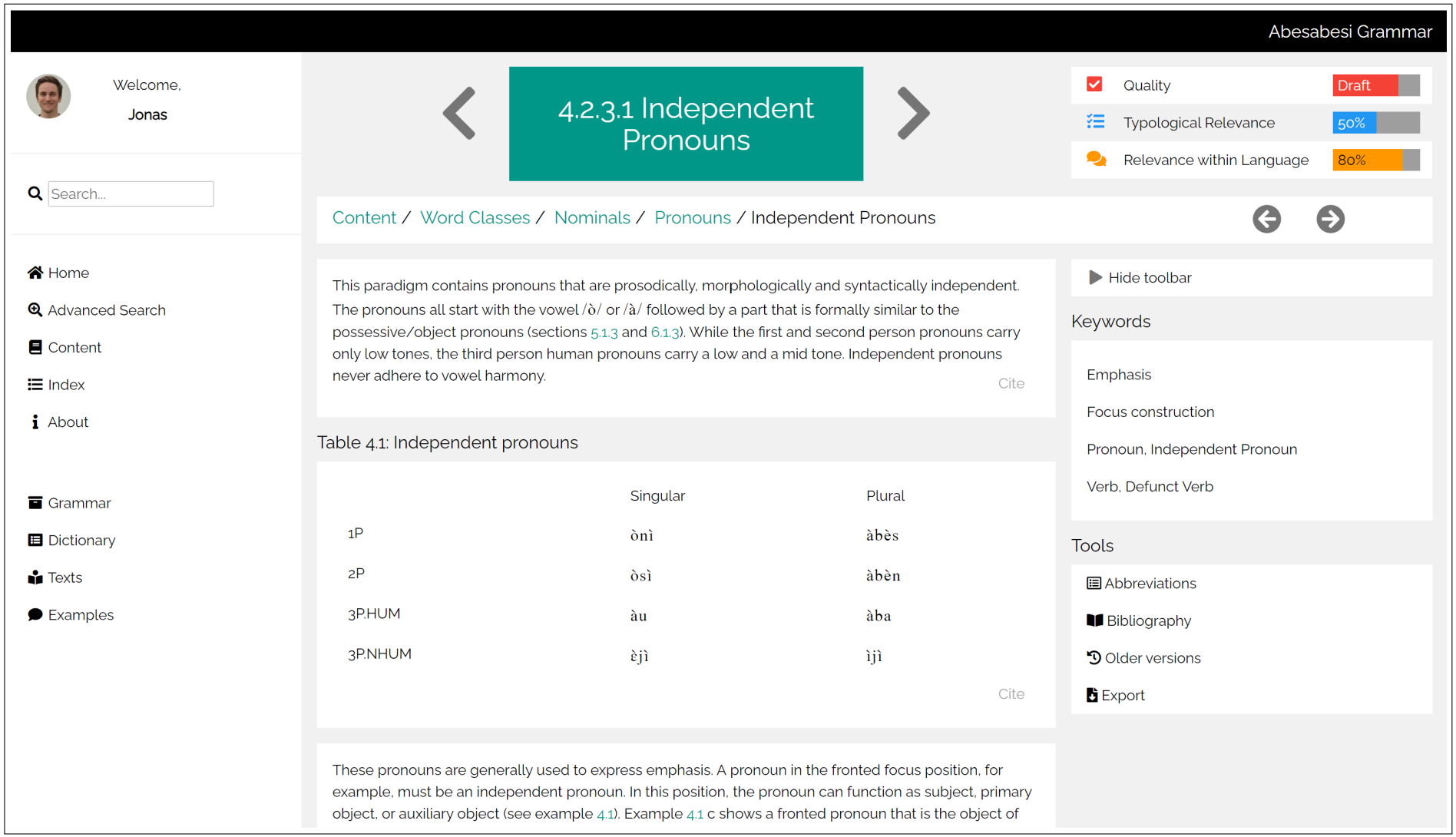


Figure 7.3: Grammar layout

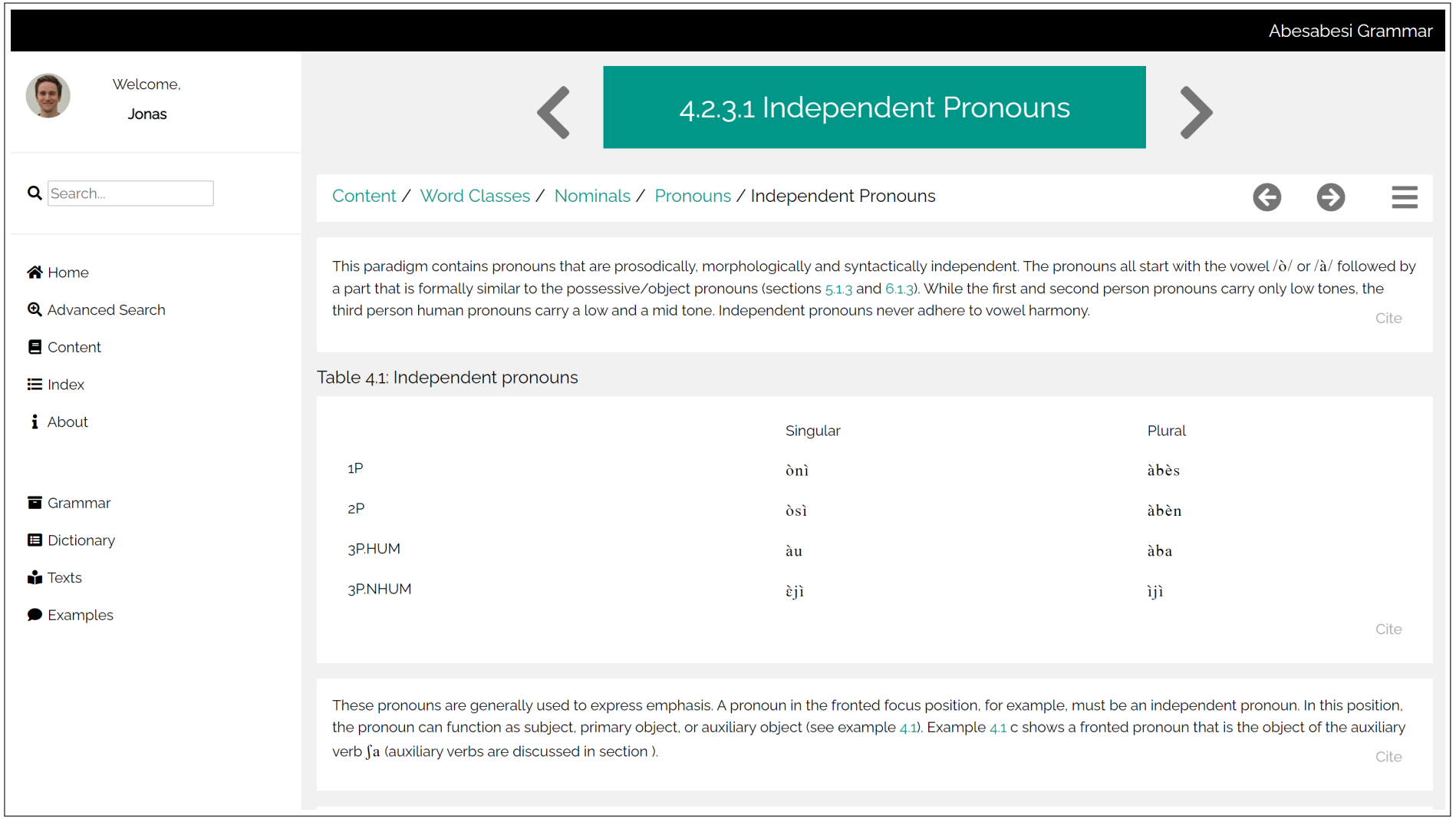


Figure 7.4: Grammar layout with hidden toolbar

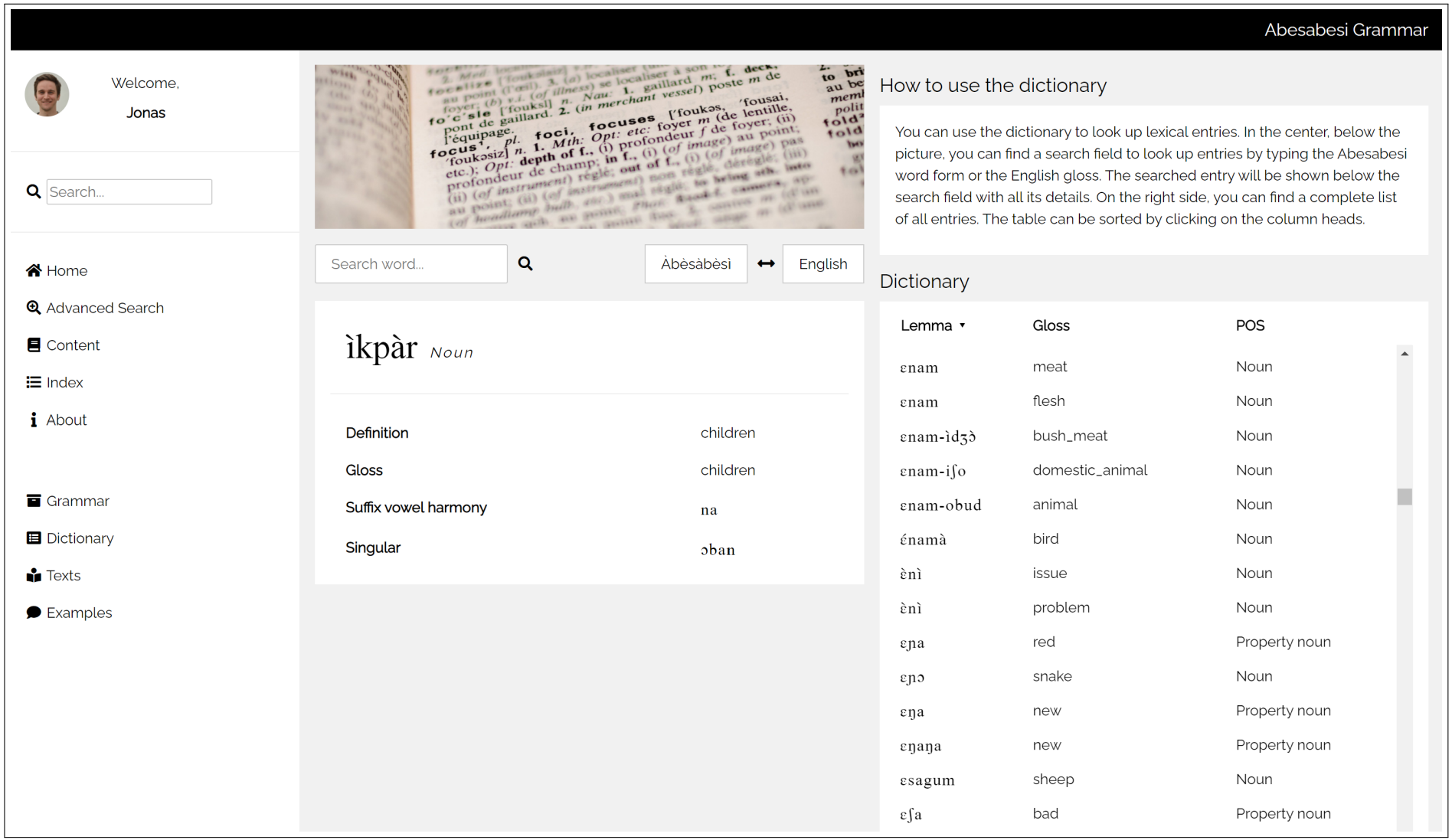


Figure 7.5: Dictionary layout

7.3.2 Grammar entries

There are various options for how to display the grammar entries, but the important factors in order to make the reading experience efficient and easy are clarity, simplicity, and familiarity. A reader will find information easier if the grammar entry's layout is clear and simple and has a structure that they are already familiar with. A familiar layout of semi-structured text is the layout of wikipedia pages. Wikipedia and other wikis usually have the heading on top, a main area where the text is displayed, and a number of boxes at the side where structured information is stored. The grammar layout in the Abesabesi GET follows similar principles (Figure 7.3). The metadata and internal navigation tools are situated on the top. Further tools are placed in the right navigation bar that can be hidden and the main content is located in the center of the page.

The heading indicates the topic and is possibly the most important piece of metadata. Therefore, it is placed in a central position on top of the page. It is surrounded by right and left arrows to get to previous and following pages. Super pages can easily be indicated by a so-called "bread-crumb" navigation (Figure 7.6). This layout device is known from other websites and indicates the position inside a hierarchy. The current page's heading is indicated on the right side and is preceded by a slash and the super-page. The super-page is also preceded by a slash and its own super-page, and so on until the top-level page is reached. Each super-page directs to the respective grammar entry.

Figure 7.6: Bread-crumb navigation

Sub-entries can be indicated by a list of their headings that direct to the respective grammar entry. The Abesabesi GET presents this list below the main content, as grammars are recommended to be read top-down. Figure 7.7 displays a (shortened) grammar entry and a list of sub-sections below. Besides the previous and following arrows to get to the previous or following entry of the path, a second set of arrows might be required to get to the last visited page. This is helpful if a cross-reference or another hyperlink has been used and the reader wants to go back to the original path. These arrows are situated to the right of the bread-crumb navigation.

Paper grammars do not display indexes within the grammar entries. If displayed as keywords at a central spot of the page, however, they can indicate whether the entry is relevant for the reader by pointing out the topics. Other metadata tags, such as those for quality assessment and relative importance (Section 5.2.1) should also be displayed at a central spot as they convey relevance before the reader starts reading. The Abesabesi GET positions them in the right toolbar. At the bottom of this toolbar, additional tools are listed. These are access to the databases that are not situated in the central navigation bar ("Abbreviations" and "Bibliography"), the version history ("Older versions"), and the output ("Export").

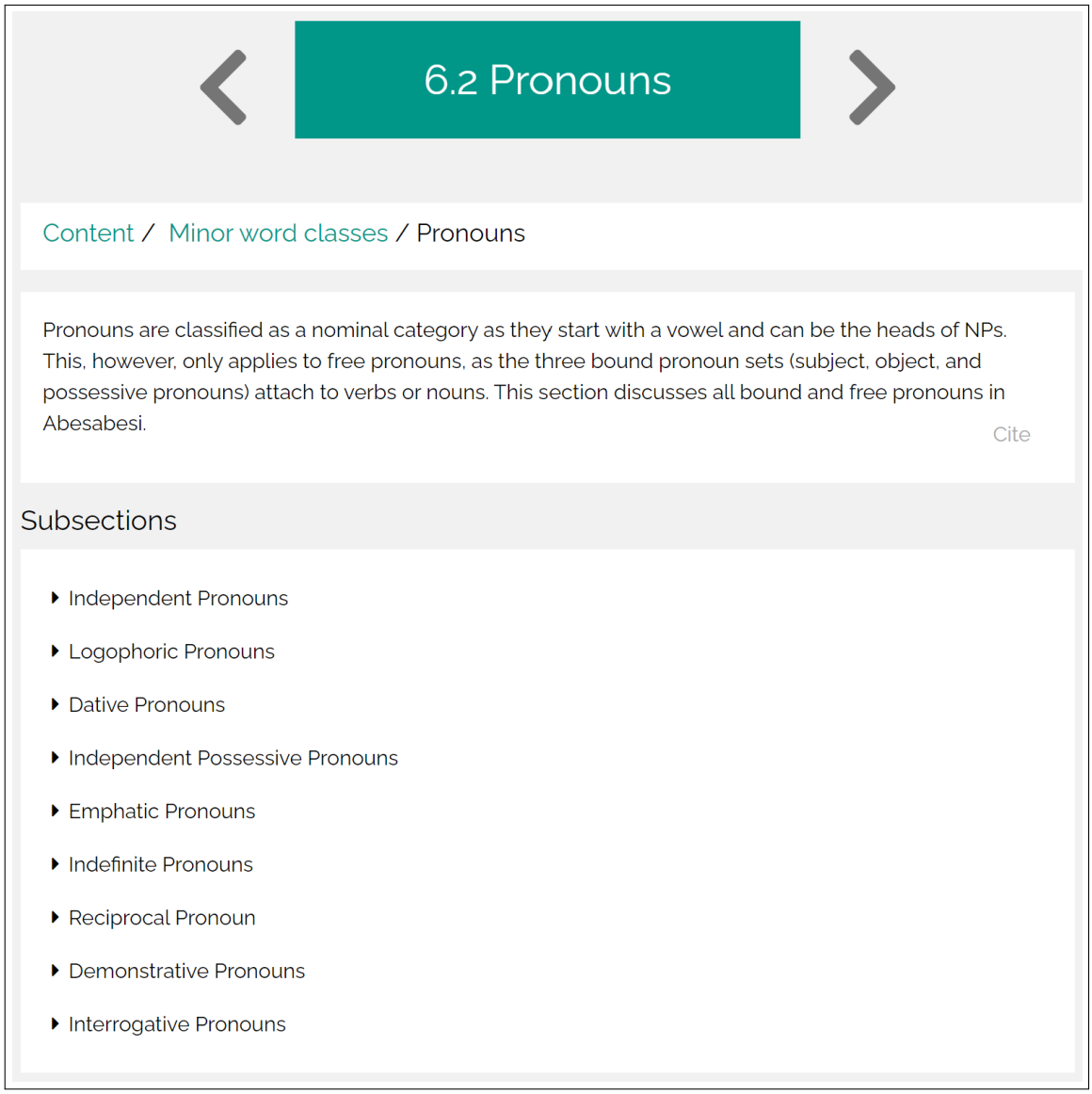


Figure 7.7: A list of subsections below a grammar entry

The main content situated at the center of the grammar entry contains all elements of the micro- and nanostructure, except for metadata. Each grammar brick (element of the microstructure) is displayed in a white box in the order of encoding. The length of a grammar entry page therefore depends on the number and content of grammar bricks. Some of the grammar bricks, such as tables, lists, or figures, can have headings that are placed on top of their respective box. Figure 7.9 shows the first grammar bricks of the grammar entry “4.2.3.1 Independent Pronouns”.

The looks of the grammar brick boxes depend on the element type. For descriptive prose, each paragraph gets its own box. While plain text within the paragraphs receives a uniform and simple format, other nano-elements such as references, linguistic units, terms, and abbreviations require special formatting.

References should be visible as references to reveal their functionality as a hyperlink. This is typically done by using another color or underlining the reference. Terms can also be seen as references because they direct the reader to the terminology entry with a definition and all pages where the term is used or explained. In the Abesabesi GET, references and terms are both formatted in a specific color to indicate they can be clicked. Although they are distinct elements and link to different parts of the website, they are formatted identically. This is to avoid too much confusion and to maintain a clear and simple structure. The formatting thus only indicates that the element can be clicked. Abbreviations can also link to the abbreviations list, but a simpler way to not disrupt the reading flow is to display the full term as a tooltip when hovering over the abbreviation. This may, however, be a problem if touch-controlled devices are used.

Linguistic units require special formatting in order to help the reader distinguish the target language from the metalanguage. They may receive additional functionalities as proposed by Drude (2012, p. 173). The italic style is traditionally used to indicate text in the described language. Additional functionalities could be triggered by a mouse click or by hovering over the linguistic unit. These include playing a recording or linking to a dictionary entry. These functionalities should still be optional, as the linking of recordings, even if semi-automated, is quite time-consuming for the author and might not be worth the effort. In the Abesabesi GET, linguistic units are all displayed in a separate font. An SIL font is used that contains all IPA symbols and is visually distinct from the default font of the grammar. In some cases, these linguistic units are accompanied by a translation which is displayed in the default font but enclosed in inverted commas.

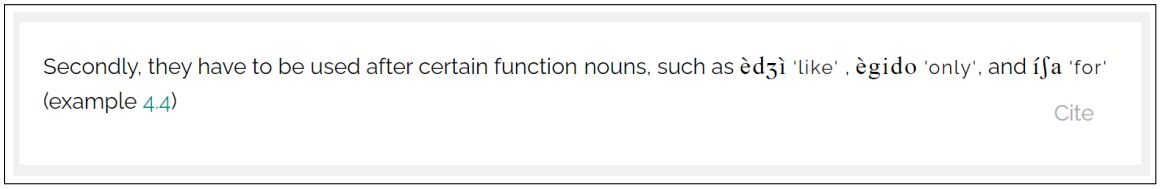


Figure 7.8: Formatting of nano-elements

This paradigm contains pronouns that are prosodically, morphologically and syntactically independent. The pronouns all start with the vowel /ò/ or /à/ followed by a part that is formally similar to the possessive/object pronouns (sections 5.1.3 and 6.1.3). While the first and second person pronouns carry only low tones, the third person human pronouns carry a low and a mid tone. Independent pronouns never adhere to vowel harmony.

Cite

Table 4.1: Independent pronouns

	Singular	Plural
1P	òní	àbès
2P	òsì	àbèn
3P.HUM	àu	àba
3P.NHUM	èjì	ìjì

Cite

These pronouns are generally used to express emphasis. A pronoun in the fronted focus position, for example, must be an independent pronoun. In this position, the pronoun can function as subject, primary object, or auxiliary object (see example 4.1). Example 4.1 c shows a fronted pronoun that is the object of the auxiliary verb ʃa (auxiliary verbs are discussed in section).

Cite

4.1 a

òní gí ʃag atɔ

1 SG.IDP FOC sweep ground

'It was me who swept the floor.' (ibe346-00.160)

more ...

b

òní gí bà kud dʒo

1 SG.IDP FOC 3 PL.HUM S do_first give_birth

'It is me they gave birth to first.' (ibe065-00.066)

more ...

Figure 7.9: Grammar bricks in a grammar entry

Lists and tables are common formats and do not require formatting different to that in books. However, it should be noted that nano-elements like terms or abbreviations do not only occur in prose but also in tables and lists. In tables and lists, they should therefore be formatted like in prose. Figures have to be saved as digital image files. All types of structured description should have a heading and a numerical identifier in order to be citable.

As discussed in Section 5.3.6, the citability of a DRG's smaller units is not obvious, as pages do not exist. Some grammar entries might be rather large and finding the information is not always easy. Instead, all grammar brick boxes have an unobtrusive "Cite" button at the bottom right corner that can be clicked to get an identifier of the grammar brick to cite. However, finding a cited grammar brick also needs to be facilitated. Therefore, the search tools also need to be able to provide a way to find elements by identifier. This feature has not been implemented.

7.3.2.1 Exemplars

The layout of examples is also the task of the GET. Examples within the prose are treated as linguistic units (Section 6.3.2.2). Their layout has therefore been discussed earlier in this section. They might also appear in lists or tables where they should be formatted like within the prose.

Interlinear examples, however, require a different formatting, as they occur in the form of IGT. Weber (2005a, pp. 458-459) mentions a few ideas for the display of examples in DRGs. The following list summarizes some of his ideas and elaborates on them.

1. Inspect the context: The context of an utterance could be displayed if wished by the reader. This would require a connection to the base text from which the utterance has been taken.

2. Toggle on/off parts: Certain parts, such as tiers in the IGT, could be toggled off or on, as the reader might not need all tiers. This is particularly important for examples with multiple tiers such as those advocated by the concept of advanced glossing (Drude, 2003).

3. Buttons and hot zones: Buttons could be used to display additional information such as the recording's metadata. Certain elements of the examples could also reveal information once they are clicked or hovered over.

4. Enhanced focus mechanisms: Besides the traditional focus mechanisms such as underlining and bold or italic fonts, enhanced focus mechanisms such as coloring and blinking could be used to guide the reader's attention.

As a default way of displaying exemplar data, I suggest the traditional three tier structure including the utterance be broken into morphemes, the morpheme glossing, and the translation. However, authors might want to have an example that deviates from the default layout. As discussed in Section 6.5.5.3, layout choices and highlighting can

be encoded together with the IGT. The application thus needs to adjust the layout to the author's choices. This concerns the number of tiers, visible and hidden tiers, alignment and highlighting. Visible tiers are displayed already and hidden tiers can be toggled on if desired by the reader (Weber's Number 2). In the Abesabesi GET, the non-default tier structure has not been implemented. Examples are always displayed in the default three-tier structure with word-alignment. However, the implementation of other format choices is feasible with more available time and will be discussed in Section 7.4. This also includes highlighting (Weber's Number 4).

Linguistic units in the exemplars can be linked to their respective lexical entries in the lexical database (Section 6.5.3.3). This way, the reader can look up more information about other meanings, pronunciation, or grammatical categories of the morpheme. Like in the descriptive prose, these linguistic units receive a hyperlink that guides the reader to the lexical entry.

The Abesabesi GET provides a couple of buttons along with the exemplars to give further information about the example (Weber's Number 3). As too many buttons under the example might distract from the example itself, only one button with the title "more..." is displayed that reveals all other options when clicked (Figure 7.10).

The first two of these hidden buttons are titled "Audio" and "Video". They reveal the respective extract of an audio or video recording that contains the depicted utterance. While in the Abesabesi GET, the audio and video recordings are stored locally, a GET could also be connected to a deposit in a language archive (Section 7.4.2). This way, the archived recordings are stored and maintained in one place, while the GET merely accesses extracts of them. If a time frame is indicated, the appropriate excerpt from the

Figure 7.10: An exemplar and context buttons

recording can be played in order to listen to the utterance in the example.

Likewise, another button is provided to display the attributed metadata. In the Abesabesi GET, this is only a mockup button, but metadata could also be retrieved via a link to the archive recording. To avoid an excessive amount of metadata, only relevant metadata, such as date, location, genre, description of the recording, and information about the speaker could be displayed. An additional button inside the "metadata" button in the Abesabesi GET would direct the reader to the archive deposit, where all remaining metadata can be accessed.

The context of a recording can be accessed if the example is linked to the text it is a part of, as described in Section 6.5.5.2. The button "Text Context" leads the reader to the right utterance within the text of the text database (Section 7.3.3.1, Weber's Number 1).

As examples can be tagged with keywords to find other examples illustrating the same phenomenon, another button reveals all keywords used as indexes for the respective example. When clicked, the button leads the reader to the term in the terminology database, where other linked examples or sections can be found.

7.3.3 Access to the databases

The Abesabesi GET grants access to the databases by providing one separate access page for each database. The only exception is the teminology database, whose content is displayed in the index navigation tool (Section 7.3.4.2).

7.3.3.1 Text database

The text database contains texts in the target language that are ideally glossed and translated. Figure 7.11 shows the text database access page in the Abesabesi GET. It uses a variant of the grammar entry layout presented in Section 7.3.1 and renders every text as one hypertext page. Similar to grammar entries, internal navigation can be managed through the arrows next to the title of a text. Additionally, a list of all texts in the right navigation bar grants direct access to each text.

The Abesabesi GET provides buttons for recordings and metadata in the right navigation bar. Additionally, it provides a direct link to the session in the archive deposit. In Figure 7.11, the video button has been clicked and a video of the entire text is presented below the button. In this GET, the example recording is stored locally and no archive connection has been established. However, Section 7.4 will discuss how an archive connection could be established. As discussed in Section 6.5.5, texts in the text database could be linked to recordings and metadata in an archive deposit. This way, the texts page in a GET could retrieve the metadata and embed an audio or video file without storing it in the application itself.

The display of the IGT in the center of the webpage is similar to that of examples on the example page and in grammar entries. Each segment is presented in a separate box with a segment identifier and the three default tiers.

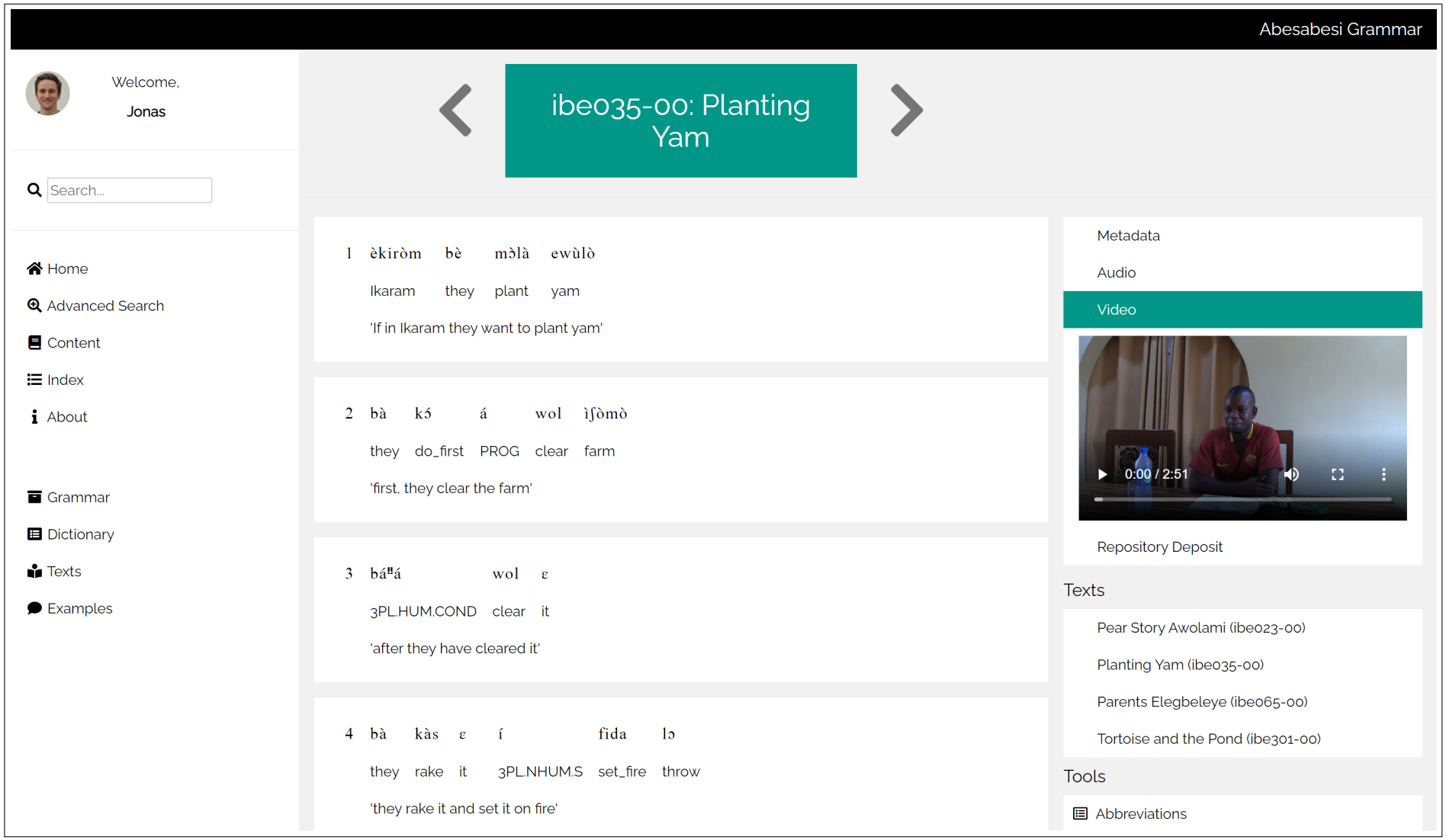


Figure 7.11: Texts Page

7.3.3.2 Example database

Providing access to the example database can be useful to the reader if the examples are tagged with keywords. This way, readers can look up other example sentences that contain a certain phenomenon. The index could be used not only to point to grammar entries but also to examples (Section 7.3.3.4). The example database could then contain even more examples as there are exemplars in the grammatical description. This is, however, only beneficial to the reader if all examples are tagged with indexes.

The Abesabesi GET has one hypertext page dedicated to all examples, using the homepage layout. Each example is presented in a separate box. This example box is very similar to that of exemplars in grammar entries (Section 7.3.2.1). The only differences are that the segment identifier is used as a heading instead of being printed after the translation and that an additional button "Used in Section" button can be clicked to see all links to instances where the example is used as an exemplar in a grammar entry. Figure 7.12 shows two examples on the examples page. The "more..." button is clicked and reveals the buttons for audio, video, the text context (to which text the utterance belongs), metadata, keywords, and sections (where the example is used as an exemplar in a grammar entry).

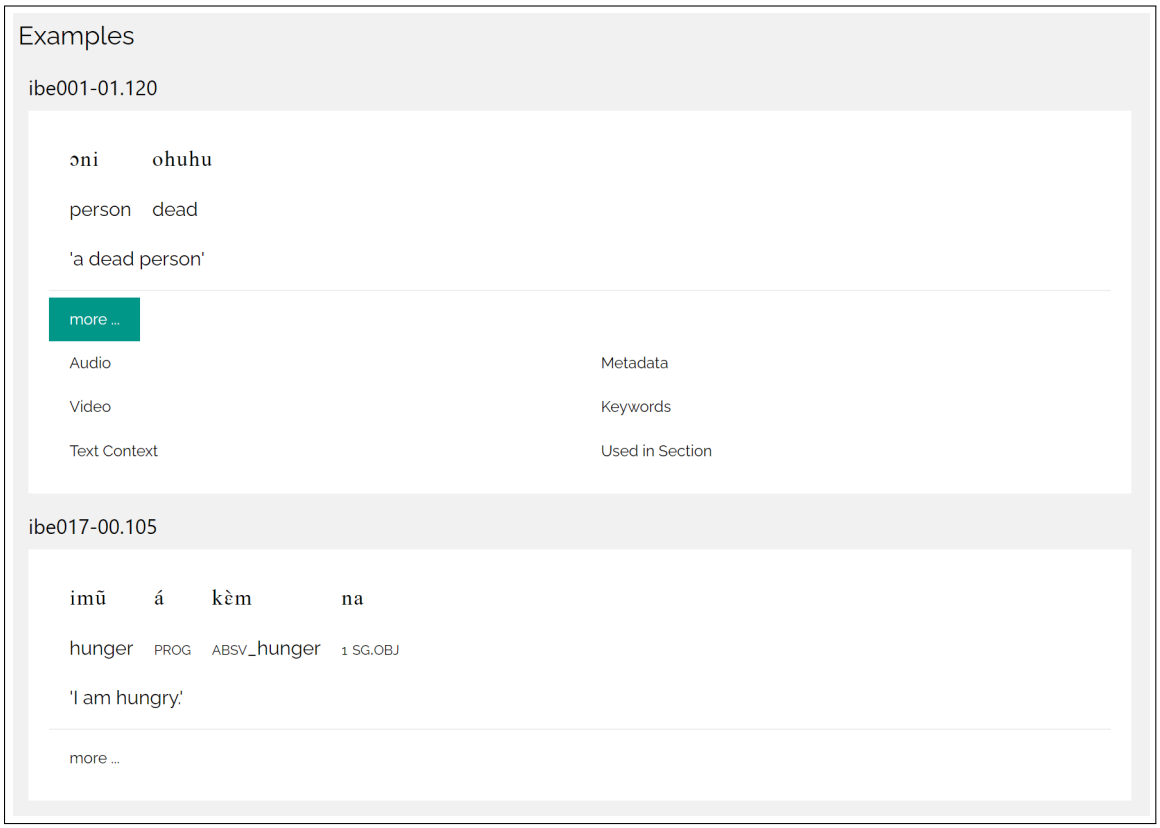


Figure 7.12: Examples in the examples page

7.3.3.3 Lexical database

The dictionary page with access to the lexical database is useful to the reader for looking up lexical entries that are used within the grammatical description. A lexeme might be used in the prose – for instance if it is part of a grammatical phenomenon, or in an example. Linguistic units can be linked to their lexical entry if encoded in the grammar documents (Section 6.5.3.3). There are various options for the structure of the dictionary within the GET. As digital dictionaries are widely used, their structure and layout is not discussed in detail. Important for the use within a DRG is only that the full entry is shown with all fields that are important to the grammatical description of the language and that lexical entries can be searched.

The Abesabesi GET uses the dictionary layout (Figure 7.5). It features a search form on top where English to Abesabesi and Abesabesi to English can be searched. The selected lexical entry is presented at the center with a definition and grammatical information. A complete list of the lexical database with sortable columns is given at the right. It contains a column for the lemma, one for the gloss, and one for the part of speech. A URL parameter "entry" is used to display the selected lexical entry in the center. The parameter is retrieved by a function in the `dictionary.xql` file and searches for the details of an entry in the lexical database. The value of the "entry" paradigm is the identifier of the selected entry. In order to look up the entry *ebo* 'dog', the URL `https://abesabesi.cceh.uni-koeln.de/dictionary.html?entry=ebo_a2d7fd54-a7c4-4fe0-b7c0-7311348a90c5` is used where `ebo_a2d7fd54-a7c4-4fe0-b7c0-7311348a90c5` is the identifier of the entry.

7.3.3.4 Terminology database

The Abesabesi GET displays the content of the terminology database in the index navigation tool. As the index already lists all terms used within the grammatical descriptions and used as indexes for grammar entries or examples, no separate access page for the terminology database is needed. Therefore, the index is a nested list of all terms displayed in the hierarchical structure of an ontology (Figure 7.13). It is discussed in detail in Section 7.3.4.2.

7.3.3.5 Abbreviation database

Abbreviations, as discussed in Section 6.5.4.4, could also be merged with the terminology database, as they usually are abbreviated terms. In this case, the access page to the terminology database should also contain the abbreviations. In the Abesabesi GET, the abbreviations are kept in a separate database and therefore have a separate access page. The page has the homepage layout and merely contains a list of abbreviated terms together with their expanded equivalent. This access page is mainly for readers that cannot access the tooltip function due to the use of mobile devices without a mouse. These readers can access the list of abbreviations and look up the meaning of an abbreviation. All other readers can hover above abbreviations and see the expanded form as a tooltip.

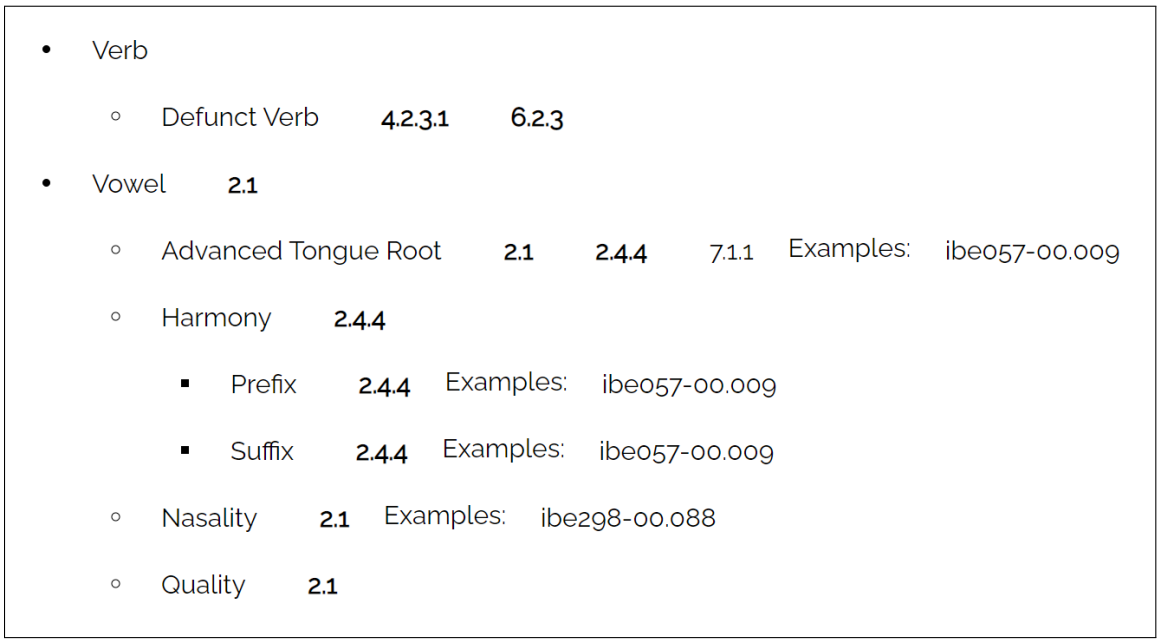


Figure 7.13: The last entries of the Abesabesi Grammar Index

7.3.3.6 Bibliography database

The last database, the bibliography database, should also be accessible to readers, as it is used to see the full reference to a document. Besides a list of all references, as known from academic documents, the Abesabesi GET displays a specific reference whenever a citation within the grammatical description is clicked. Like the dictionary and grammar entry pages, the specific entries are displayed through a URL parameter.

In order to adhere to each reader's needs, the formatting style of the reference could be selectable and an output could be given in various bibliography formats such as BibTeX (file extension `.bib`) or the Research Information System Format (file extension `.ris`). This has not been implemented in the Abesabesi GET.

7.3.4 Navigation tools

The navigation tools a GET could offer have already been presented in Section 5.4.2.4: a TOC, an index, search forms, version history, and layer selection. This section will now discuss how each of them was implemented in the Abesabesi GET.

7.3.4.1 TOC

The TOC implemented in the Abesabesi GET follows the traditional model: A standard view of a nested list of the grammar entry headings. As proposed in Section 5.4.2.4, the deeper levels are un-/foldable. Figure 7.14 displays the TOC where Chapter 5 is unfolded.

Other approaches have not been implemented as they require special encoding of

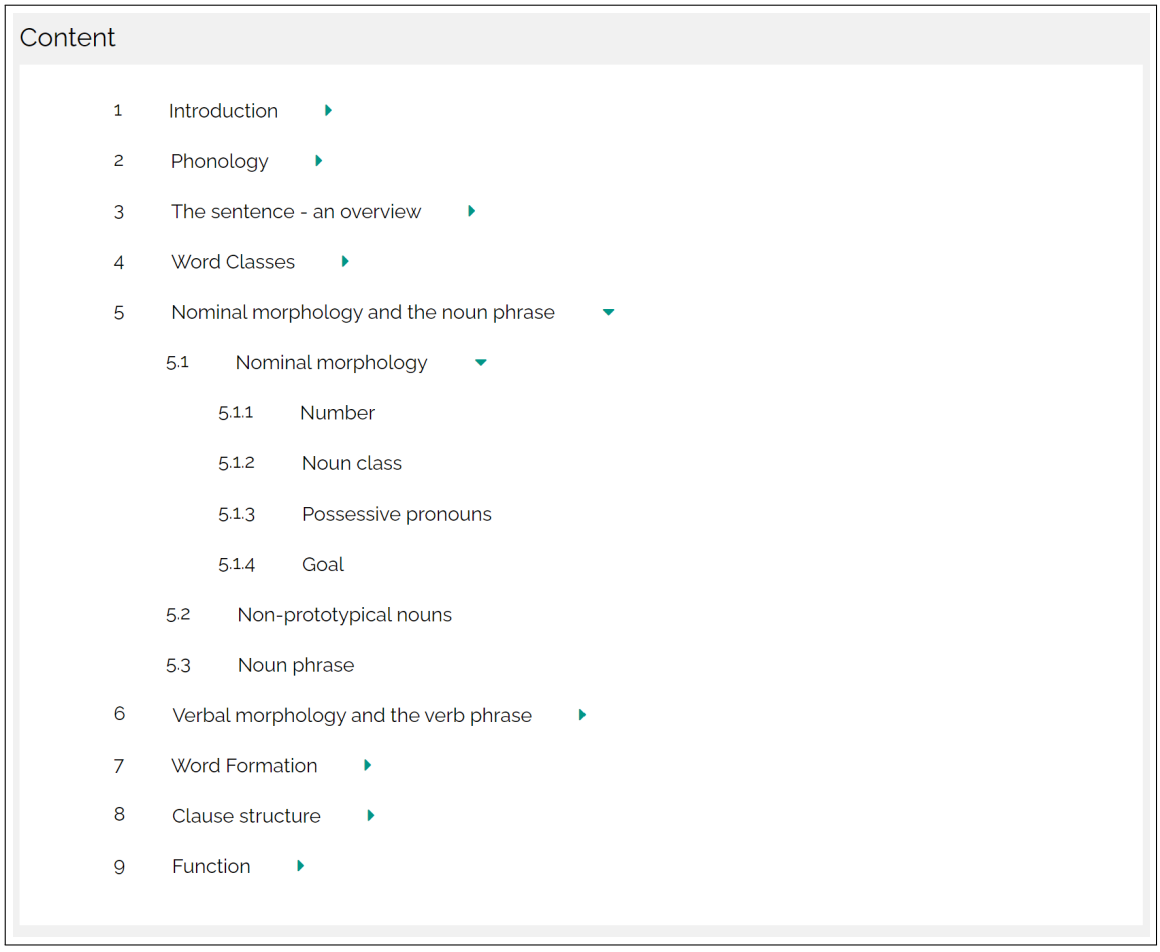


Figure 7.14: The TOC in the Abesabesi GET

the grammatical description (the form-function approach, the dynamic path approach, and the marking of read entries). How their implementation could look and whether the different approaches could be combined will be discussed in Section 7.4.

The TOC is generated by iterating through the grammar entries and accessing their headings. Each time a nested grammar is found, a nested list item is created in the TOC.

7.3.4.2 Index

The index of the Abesabesi GET contains all terms of the terminology database that are used within the descriptive prose or as indexes for grammar entries and examples. It therefore also functions as an access page for the terminology database.

This merge is basically a merge of the two traditional navigation tools index and glossary. It was, however, facilitated by the fact that only one ontology was used. This way, only one ontology needs to be presented as a nested list that is sorted alphabetically (Figure 7.13). Using terms from more than one ontology, on the other hand, would create a couple of challenges.

One approach would be to present all ontologies and display them as separate nested lists. The first problem arising with this approach is that this kind of index would not be very handy, as readers would need to search a certain index in multiple lists instead of in one. Secondly, authors might only use several terms from one ontology and leave the rest of the ontology unused. Should the index thus display the whole ontology in order to demonstrate the terms' position within the hierarchical structure? If so, the unused terms of an ontology might be distracting for the reader and could be redundant as some terms might be part of multiple ontologies.

The second approach would be to separate the index and the terminology access page. The index would then only be used as a navigation tool and contain the terms used as indexes. The terminology access page, on the other hand, would function as a glossary and display the ontologies with definitions of the terms. In this approach, indexes should not only refer to the respective grammar entries and examples, but also link to their entry in the terminology access page. This way, the indexes' definitions can be looked up. If the index should have some nesting, this would need to be encoded somehow. The Abesabesi GET nests the indexes according to the hierarchical structure of the terminology database.

The merge of the index and terminology databases was also facilitated by the fact that the used ontology did not have too many levels. As an index is sorted alphabetically and might have one or two levels, ontologies with many levels might not be easy to use as an index. They might only have a couple of terms in the first level and a term needs to be found inside the structure of the ontology. This might be hard if one is not familiar with an ontology's structure.

The Abesabesi GET accesses all terms of the terminology database and displays them in a nested list. For each term, it trawls through the grammar entries and selects those that contain the term. It prints the identifiers of grammar entries that contain the respective term as an index in boldface and those that merely contain the term in the grammar bricks

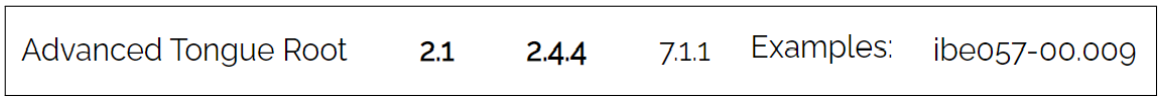


Figure 7.15: A term with identifiers of associated grammar entries and examples

in the default format. Example identifiers are displayed after that. All identifiers are rendered as buttons that can be clicked to reach the respective grammar entry or example. Figure 7.15 shows the term “Advanced Tongue Root” in the Abesabesi GET. Grammar entries 2.1 and 2.4.4 use the term as an index (boldface) and Entry 7.1.1 contains the term within the grammar bricks. Example ibe057-00.009 uses the term as an index.

7.3.4.3 Search forms

Search forms are navigation tools that allow readers to find specific keywords quickly. As digital documents can be scanned by computers in less than a second, these navigation tools constitute one of the advantages of DRGs. A simple full text search form is located in the main navigation bar of the Abesabesi GET (Figure 7.16). It is accessible from every page of the GET and enables the reader to type in words that are compared with the content of the grammar documents.

The simple full text search leads the reader to a result page that is generated by the XQuery file `search.xql`. Within the homepage layout, it lists different kinds of search results of the query – each of them leading to a grammar entry that contains the search phrase. The search results are ordered: grammar entries that have the search phrase in the heading are listed first, those that are indexed with the search phrase second, and those that have the search phrase somewhere as plain text last. The results for the last criterion are presented as paragraphs with a link to the respective grammar entry. How results are displayed in the Abesabesi GET can be seen in Figure 7.17. A more detailed ordering of the results could be implemented to make this navigation tool more effective. For instance, the results could not only be ordered by the element the search phrase appears in (headings > indexes > terms/abbreviations > grammar bricks) but also by how often the search phrase appears in a grammar entry. Furthermore, the simple full text should also recognize identifiers in order to lead the reader directly to the identified element.

Advanced search forms have only been implemented as a mockup page in the Abesabesi GET. The button “Advanced Search” in the navigation bar (Figure 7.16) leads to a page that contains three search forms (Figure 7.18).

The first search form is titled “Search by ID”. This form helps by finding a specific grammar brick that has been cited in another text. For instance, a paper cites Table 2.4 of Version 1.2 of the Abesabesi Sketch Grammar (Figure 7.19). In order for readers to find the respective paragraph in the right version of the grammar, they need to chose

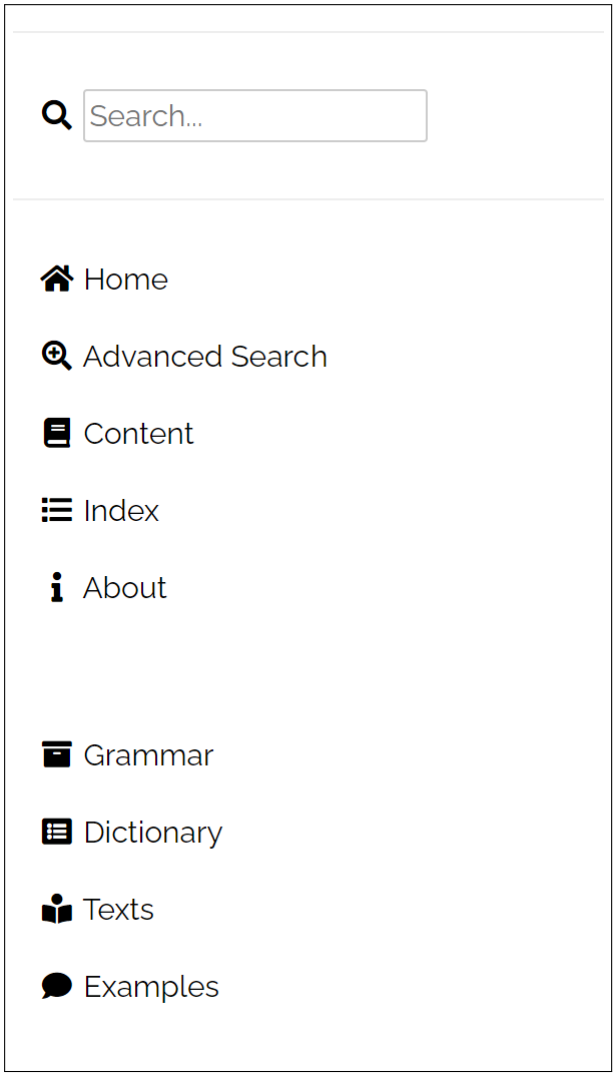


Figure 7.16: The main navigation bar with search forms

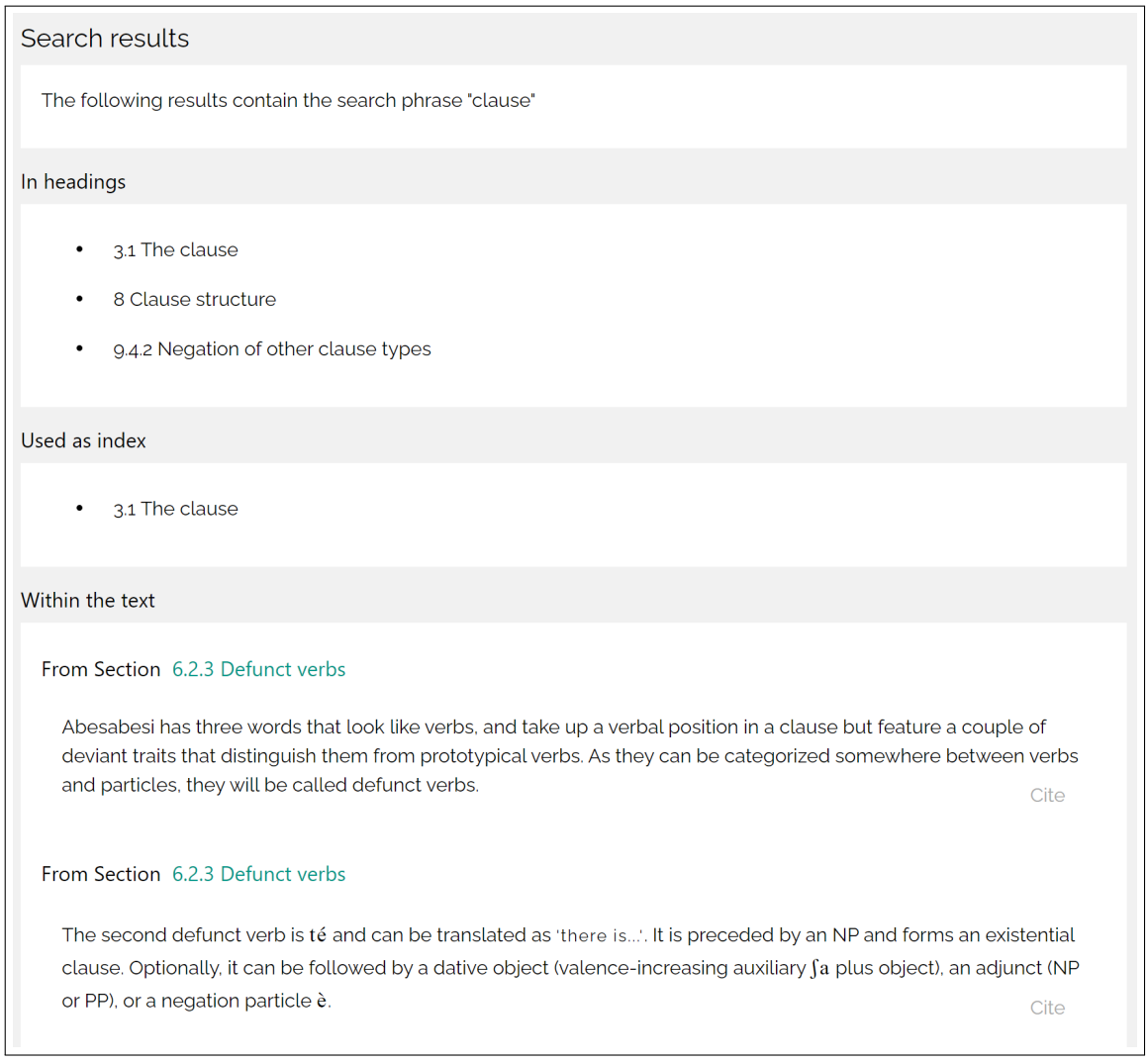


Figure 7.17: Displaying search results

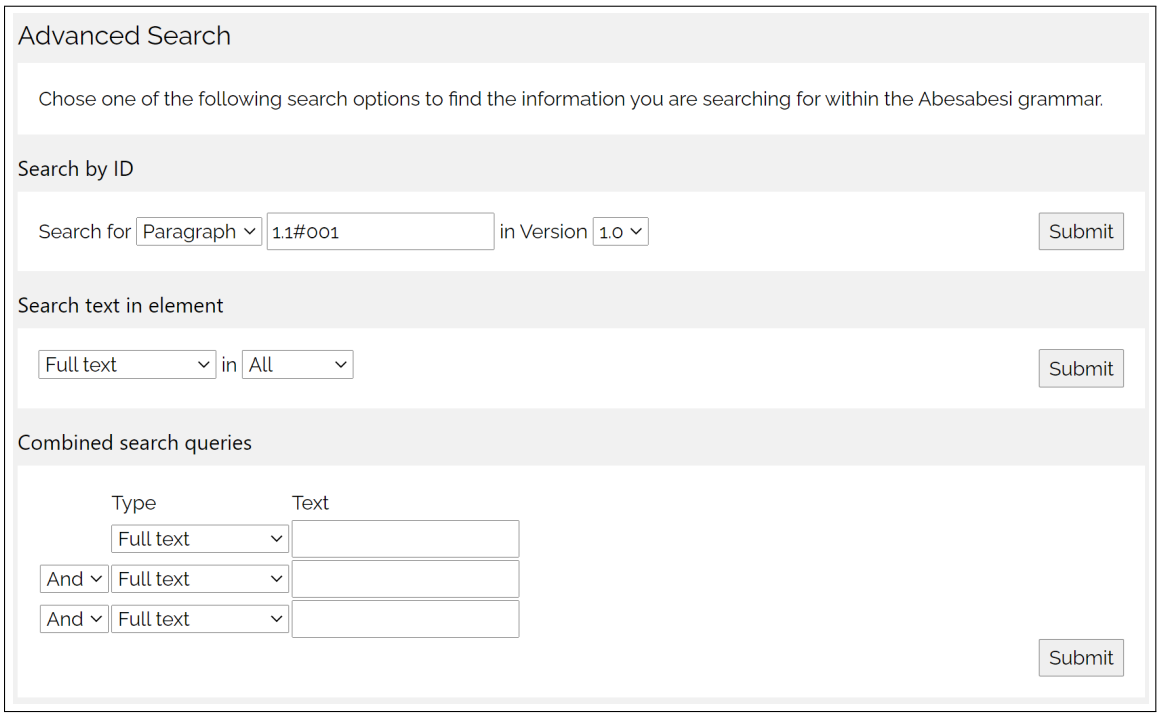


Figure 7.18: Advanced search forms

the kind of identified element (“Table”) in the dropdown box, type the identifier (“2.4”), and select the right version number (“1.2”) in the form before submitting. Then, they would be guided to the right paragraph. All identified elements could be searched with this method (grammar entries, paragraphs, figures, tables, lists, interlinear examples).

Figure 7.19: Search by ID

The second search form (“Search text in element”) searches for specific nano-elements in specific grammar bricks or grammar entries. For example, a reader could want to find all instances of the Abesabesi word *àu* ‘3SG.HUM.IDP’ within interlinear examples (Figure 7.20). They would need to select the nano-element in the first drop-down box (“Lexeme”), type the word and select the element that could contain the nano-element (“Interlinear example”). It would be helpful for readers to have an autofill function when starting to type an Abesabesi word. This way, the reader could ensure that the desired word is written correctly and has an entry in the lexical database. Possible nano-elements that can be searched with this search form are plain text, terms, abbreviations, text in Abesabesi, and different kinds of linguistic units (phones, phonemes, syllables, morphemes, etc.). An autofill function for each element with a limited set of entries

(terms, abbreviations, lexical entries) could be beneficial for the reader.

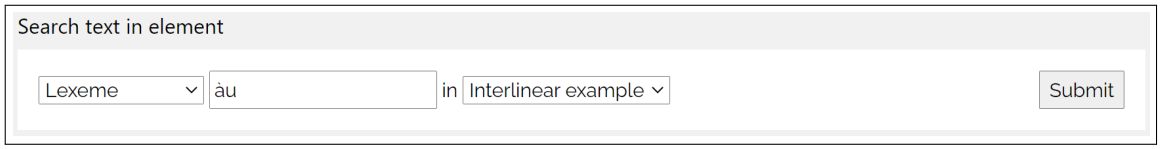


Figure 7.20: Search text in element

The third search form is for combined search queries. Its search results are all grammar entries that meet the selected criteria. Each line represents one criterion. These criteria can be combined by "and" (both criteria have to be met) or "or" (one of these criteria has to be met). For each criterion, an element type can be selected and the content of the element can be typed. For example, all grammar entries containing the Abesabesi word *ɔ́ńtɔ̀ǹ* 'anybody, nobody, whoever' and the term "negation" could be searched in order to find out whether and how indefinite pronouns are used in negation (Figure 7.21). Therefore, in the first line, "Abesabesi text" is selected and the text "ɔ́ńtɔ̀ǹ" is typed into the form. The second line contains the option "term" and the text "negation".

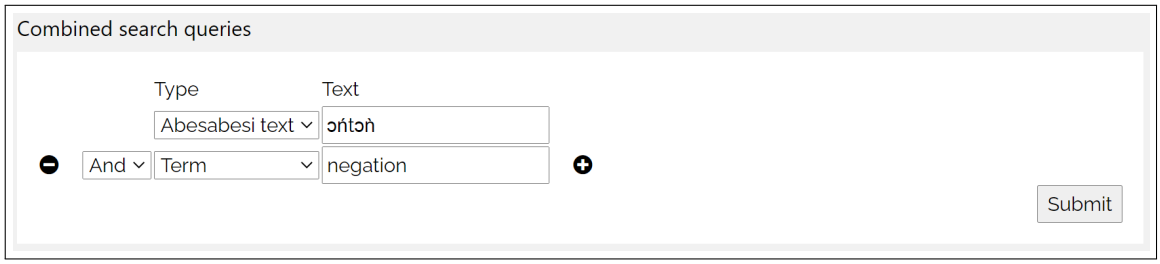


Figure 7.21: Combined search queries

All of these advanced search forms are only mockups and queries are not processed. Implementation, however, should not present a big problem, as these kinds of forms are common to many databases.

7.3.4.4 Other navigation tools

Version history and its benefits have been discussed in Section 5.3.4. The tool should enable readers to access older versions of grammar entries or databases and see which edits have been made by which author. The Abesabesi Sketch Grammar has only been published once as one version with one author. Therefore, it will not be an ideal example for version history. The tool has therefore only been implemented as a mock-up button in the right navigation bar for grammar entries and databases. The button reveals further buttons for several imaginary versions and an overview to see all page edits (Figure 7.22). What a functioning version history could look like will be discussed in Section 7.4.

The last navigation tool, layer selection, has not been implemented in the Abesabesi GET, as the grammatical description has not been encoded in layers. How it could be implemented in a GET will be discussed in Section 7.4.

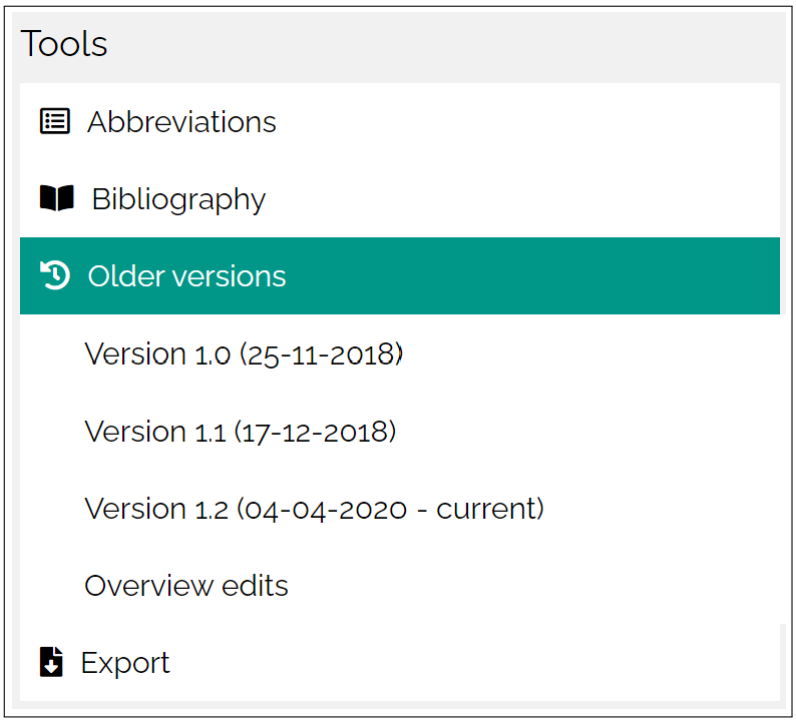


Figure 7.22: Version history mockup

7.4 Unimplemented features

This section lists those features of the outline (Section 7.1) that have not fully been implemented. Each subsection discusses one feature and explains how it could potentially be implemented and what kind of challenges the implementation would entail.

7.4.1 Login facility and offline usability

The Abesabesi GET does not distinguish users and does not allow for registration or login. As discussed in Section 5.4.2.1, a login facility could be necessary in a commercialized GET, in a GET where the data of the grammatical description should be restricted, or in a GET that stores and displays which grammar entries have already been read. Implementing a login facility would require three parts: the application logic of storing and managing user data, a registration facility, and a login facility in the GET. Registration and login facilities are known from all kinds of commercial and non-commercial websites and their implementation would not require a lot of resources per se. The application logic to store and manage user data, however, would entail a few more tasks. First, user data would need to be stored in some kind of document. This could, for instance, also be an XML or TEI document that could be structured as a database. Second, a further application module would need to manage the connection between login and registration facilities and the user database. If the access should be managed in more detail (e.g. only some users will be allowed to access all parts of the grammar), human user management

might additionally be required. Therefore, the login facility would need to lead registered administrators, such as the author, to a webpage where all users could be seen and the author could grant rights or delete accounts. As personal data is stored, it would also be important to find an ethical and legal way to treat and manage the data. In the EU, this would, for instance, fall under the General Data Protection Regulation.

There are plenty of tools nowadays to facilitate the type of offline usability that was discussed in Sections 5.2.2 and 5.4.2.1. Accessing resources that are usually accessed through the internet is enabled by the concept of caching. In this process, the browser stores the resources essential to continue running the web application locally. This concept has, for instance, successfully been implemented for the CMDI Maker[2], which enables users to manage metadata and continue their projects without internet connectivity. A general implementation of this feature would thus be possible. Some features of a GET could, however, be restricted when offline. For instance, media files could not be played if they are only accessed by the application through an application programming interface (API) (see following section).

7.4.2 Archive connections

Instead of storing video recordings and metadata locally, like in the Abesabesi GET, a GET could ideally access an archive and its deposits to play resources stored there. This would save storage space and maintain one central location where the recordings are stored.

As requirements for this archive connection, the archive needs to use persistent identifiers and the deposit needs a persistent structure in order to guarantee minimal maintenance and a stable connection. As archives are build to persist and endure, this should theoretically not pose a big problem. In practice, however, longterm stability is often not guaranteed, as language archives depend on external funding.

In order for applications to access data from an archive, they could use APIs. APIs enable applications to interact and exchange data. The Language Archive Cologne[3], for example, provides a set of APIs that allow other applications to access media or metadata[4].

Like deposits in archives, grammar authors also need to be cautious and ethical about which data is made openly accessible. As archive data is personal data, recorded individuals need to give their consent before anyone is allowed to publish the data online. Recordings in a deposit are thus restricted at times or might only be accessible with a registered account. This is also important for GETs. A GET can only access archive data that is openly accessible. If login is required to access certain data, the authentication process could also be handled through APIs.

[2]retrieved 2021-01-10, from `http://cmdi-maker.uni-koeln.de/`

[3]retrieved 2021-01-12, from `https://lac.uni-koeln.de/`

[4]The documentation can be accessed at `https://ka3.uni-koeln.de/apidoc/index` (retrieved 2021-01-12)

7.4.3 Example layouts

The Abesabesi GET only presents a default layout of the examples and does not allow for deviating example layouts and alignments, toggling on and off additional tiers, or the highlighting of certain elements. However, these features need to be implemented in a GET in order to present the example exactly how the author intended to. As discussed in Section 6.5.5.3, all of these features can be encoded in TEI-IGT. How they could be implemented will be illustrated in the following paragraphs.

Highlighting elements of an utterance can be a method to direct the reader's attention to these elements. Figure 7.23 displays the formatting of an example that has been used before to illustrate TEI-IGT. Such an example might be used in a grammatical description to illustrate how gerunds are formed with a circumfix. Therefore, the circumfix is highlighted. In this case, boldface and a contrasting color were used to highlight the two morphemes. The style of the highlighted elements needs to be noticeable in order

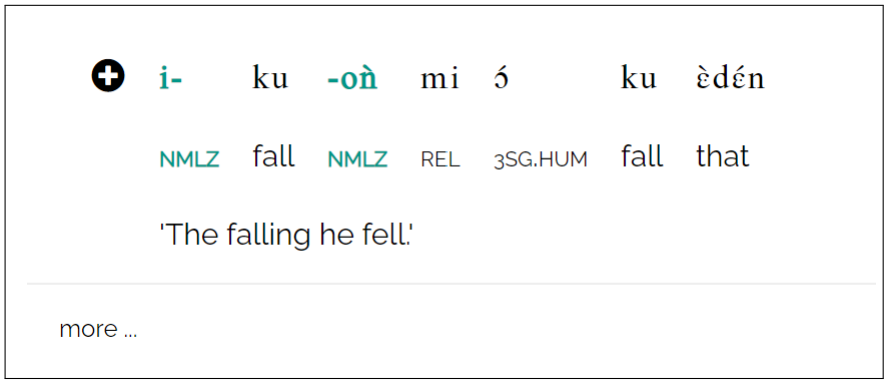


Figure 7.23: Example with morpheme alignment and highlighting

to attract the reader's attention. Furthermore, it needs to be differentiated from other kinds of styling that bear meaning. For example, the focused phrases in pragmatically marked constructions are often styled in all capitals and should not be confused with simple highlighting.

The example in Figure 7.23 also features morpheme alignment (Section 6.5.5.1) as opposed to the default word alignment that is, for example, used in the Abesabesi GET (Figure 7.12). As mentioned before, this kind of alignment does not follow the Leipzig Glossing Rules (Comrie et al., 1982) but might be preferred by some authors since the relation between morphemes and glosses is a little clearer.

Moreover, a plus icon is displayed next to the first tier. When clicked, it reveals an additional tier – the word tier (Figure 7.24). The additional tier can now be toggled off by clicking the minus icon. Providing additional tiers can be helpful for readers that like to find out more about the examples or for authors to clarify issues that are not directly important for the current argumentation. In this example, for instance, the author might want to illustrate the relativized focus construction and stress that the gerund circumfix is used in the construction. Opening the additional word tier could reveal to interested readers how the gerund looks on a word level after adding the circumfix.

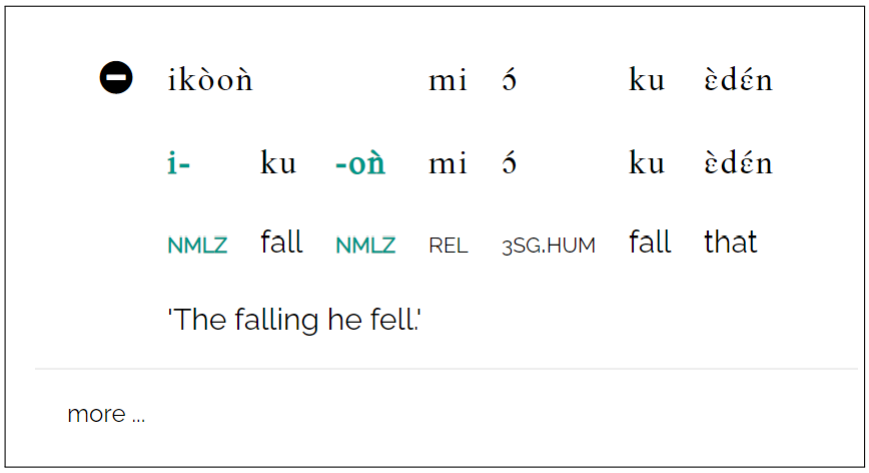


Figure 7.24: Additional tier is toggled on

7.4.4 Data output

A GET that allows for multiple kinds of data outputs facilitates further processing of the data. A raw output of the grammatical description in the TEI format, for instance, could be used by other applications to extract data. The application might aim at the comparison of certain linguistic features. Having access to the TEI data, it is now able to distinguish linguistic units, terms, and other elements, provided they are encoded correctly. Data outputs thus follow the value “manipulation” discussed in Section 5.2.9.

Access to the raw data could be given for the entire grammatical description, entire databases, or specific elements such as grammar entries or interlinear examples. The bibliography database could, for instance, be extracted as a whole or the output could only contain one specific entry, depending on what the reader needs. The data could also be transformed automatically and be provided in other formats. Readers might want to download a bibliographic entry in another data format they can directly incorporate into their preferred bibliography management software.

In order to allow other applications to access the formated data, APIs could be provided (an explanation of APIs is given in Section 7.4.2). Otherwise, a person would need to download data and ingest it into their application. This can be time-consuming and would need to be repeated anytime the grammatical description is updated or extended. The development of APIs would be additional work but should be feasible and would greatly foster the dissemination of the author’s research. However, the extraction of data by applications can also be prone to various errors. For example, faulty encoding could result in a false attribution of data. For instance, an author might use the element `<foreign>` to encode a term, only to ensure a specific styling within the GET. If this faulty encoding is now extracted by an application, the term is now interpreted as text of the target language. Another example is that comments on the certainty of a specific description might simply be ignored. A speculative statement would thus be extracted

by an application and incorporated into a database where it will be used as the base for multiple new hypotheses. This problem could be avoided if the certainty is encoded in a standardized way. This encoded certainty could now be extracted and tied to the respective data piece.

While applications would benefit from access to the data through APIs, human readers would benefit from a way to download data from the GET. The Abesabesi GET has mock-up buttons in appropriate places to download data. In the right toolbar for grammar entries, for instance, there are buttons to download the entire grammatical description or the selected grammar entry as TEI or PDF. Likewise, a GET should provide similar buttons on every database access page.

Providing linear formatted outputs, such as PDF versions of the grammatical description, could be beneficial for readers that prefer to hold paper in their hands. The aspect of "tangibility" was discussed in Section 5.3.1. A home-printed PDF might not be a worthy substitute for a grammar book, but it might be preferred by readers that would like to take hand-written notes when reading a specific grammar entry. Producing a PDF output requires a transformation from TEI to PDF. This is only possible if the grammatical description is structured linearly or if a complete path through the grammatical description is provided by the author.

7.4.5 TOC alternatives

While the Abesabesi GET follows a rather traditional approach to displaying the TOC, there are more approaches to structure a TOC in a GET. As the TOC represents a grammar's structure, following an alternative approach also affects how the grammar's structure is perceived and explored by the readers. As discussed in Section 5.3.7, a linear grammar with a static path from the beginning of a grammar to its end also requires a linear structure of the grammatical description. It contains a coherent story, an order of grammar entries, and often transitions between grammar entries to guide the reader. The traditional TOC design is aligned with this structure and represents the hierarchical but linear structure of a grammatical description.

A dynamic path as mentioned in Section 5.3.7 does not require a linear structure of the grammatical description. It only requires the attribution of grammar entries to certain domains and the encoding of each grammar entry's priority within the domain. The affiliation to a domain and the priority can both be tagged as metadata (Section 6.5.2.1). Implementing a TOC representing this kind of structure could look like in Figure 7.25. This TOC displays four sample-domains. The associated grammar entries are revealed when the triangle next to the domain name is clicked. They are ordered according to their tagged priority. A short introductory text could be reached by clicking on the domain name. Within this text, the domain could be explained and the topics that belong to the domain could be introduced. This approach would also require some changes in the internal navigation tools. If a grammar entry is read, the arrow to navigate to the next entry needs to point at the entry that has the next highest priority in the same domain.

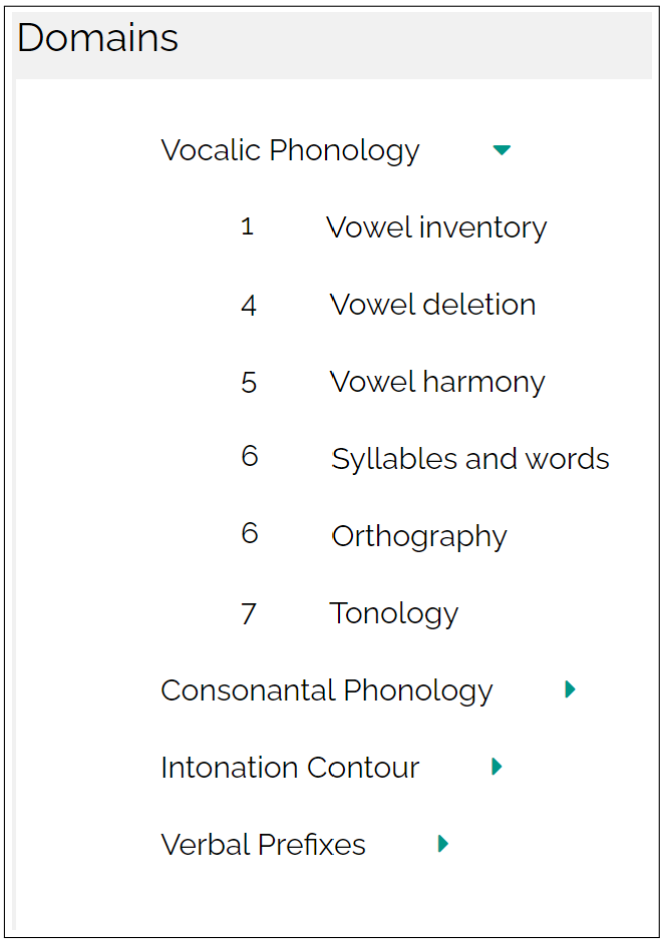


Figure 7.25: Domain-centered TOC for the dynamic path approach

The third approach is the form-meaning approach discussed in Section 6.1.3. Most grammar entries represent a form-meaning pair that either describes a form and related functions or a function and related forms. As each form is related to at least one function, these relations could also be expressed in the TOC. Figure 7.26 displays such an approach. Grammar entries representing a form-meaning pair could all have a specific icon to their right. This icon (double arrows) reveals the related functions or forms when clicked. In Figure 7.26, the icon reveals a grey box where the function of the particle *mV* is displayed. The function is represented by the heading of a grammar entry. This heading can also be clicked and will guide the reader to the grammar entry that discusses polar questions.

Theoretically, these three approaches could be combined, as a grammatical description could be encoded linearly, domains and priorities could be tagged, and form-function relations could be linked. While formally this seems to be implementable, writing such a grammar could be a rather difficult undertaking. A grammatical description that is written as a linear story could be rather confusing if displayed in a dynamic path. Transitions would need to be avoided and references that are implicit due to grammar entry adjacency would need to be expressed through cross-references. Writing such a grammar

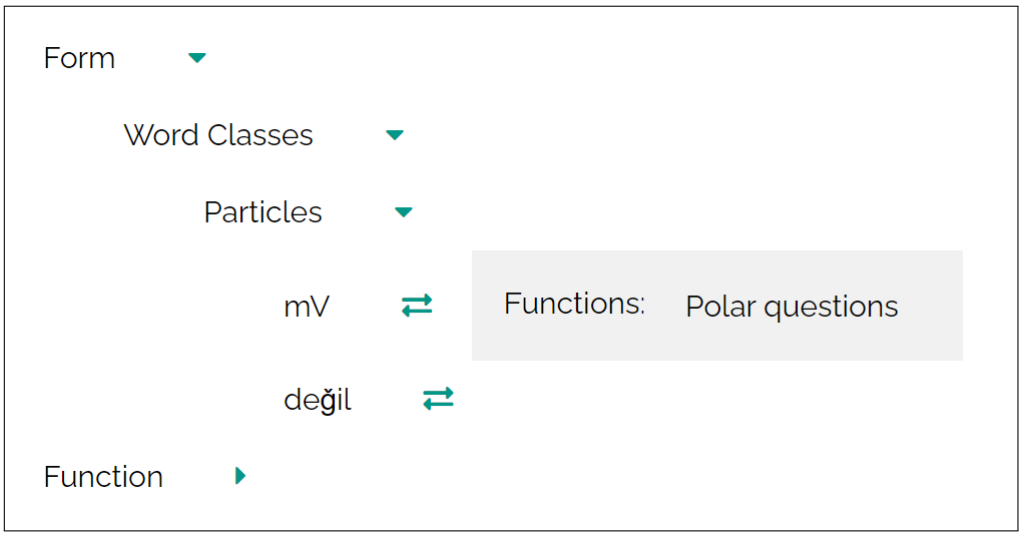


Figure 7.26: Form-meaning approach to a TOC

could end up being rather confusing and perceived as an incomplete compromise.

An additional feature of TOCs that could be implemented is to mark read entries. This method has been discussed in Section 5.3.7. Read entries could be indicated in the TOC to let the reader know which entries still need to be read in case they intend to read the whole grammar. This is especially important for a dynamic path. Readers might want to read a whole domain and would want to know when they have read the last entry attributed to a domain. Moreover, entries might be attributed to multiple domains and readers might have already read an entry when exploring another domain. Marking read entries as read could be done by styling them differently. This method is a familiar feature of websites where links that have already been clicked are styled in a different color and underlined. However, it usually only works for one session at a specific website. In order for a GET to memorize which grammar entries have already been read, the reader needs to be recognized. A solution to this is user management (Section 7.4.1). If users are logged in, read pages can be stored together with the user's personal data.

7.4.6 Version history

In terms of a version-history tool, the Abesabesi GET only provides an "Older versions" button accessible from each grammar entry (Figure 7.22). The button reveals mockup links to all existing versions and a link to an overview of the different edits. The links to existing versions are supposed to allow readers to access an older version directly. This might be the case if they want to follow a citation in another publication. These buttons thus act as a layer selector, since once a version is selected, the grammar entries and the databases are displayed according to the selected version. However, some readers might not recognize that they have switched to another version or forget they are currently reading an old version. Therefore, it would be helpful to have a reminder of which

version is currently displayed. This could be, for example, a note in the toolbar that says "Currently accessed version: 1.2". This is particularly important if the accessed version is not the most current version. In this case, the note could be printed in another color or be accompanied by a small warning sign (Figure 7.27).

An additional page containing an overview of all edits of a particular grammar entry could list all version numbers, the authors that published the respective version, and a summary of the edits made. An example for this type of version history is that of Wikipedia articles. As multiple authors can edit an article, it is important to retrace which author edited what. Each version can be compared to another by selecting two versions and clicking a "compare" button. The page now displays all elements that differ from each other in the respective versions (Figure 7.28)[5]. While the older version is displayed on the left, the newer version and the edits are displayed on the right. This way, a reader can retrace what exactly has been changed from one version to another.

[5]The figure is an extract from the version history of the Wikipedia entry "Akpes language" (retrieved 2021-01-12, from `https://en.wikipedia.org/w/index.php?title=Akpes_language&type=revision&diff=891946710&oldid=891946230`

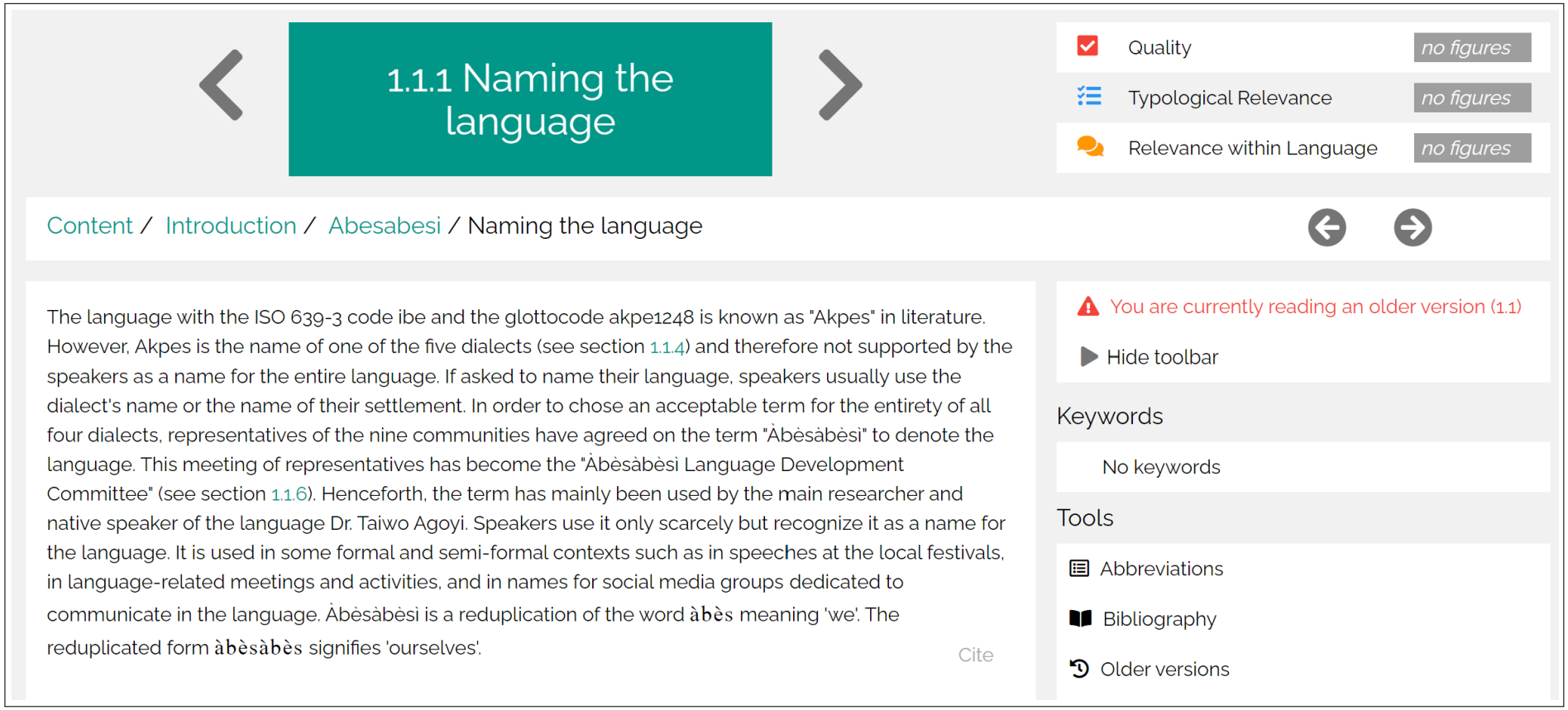
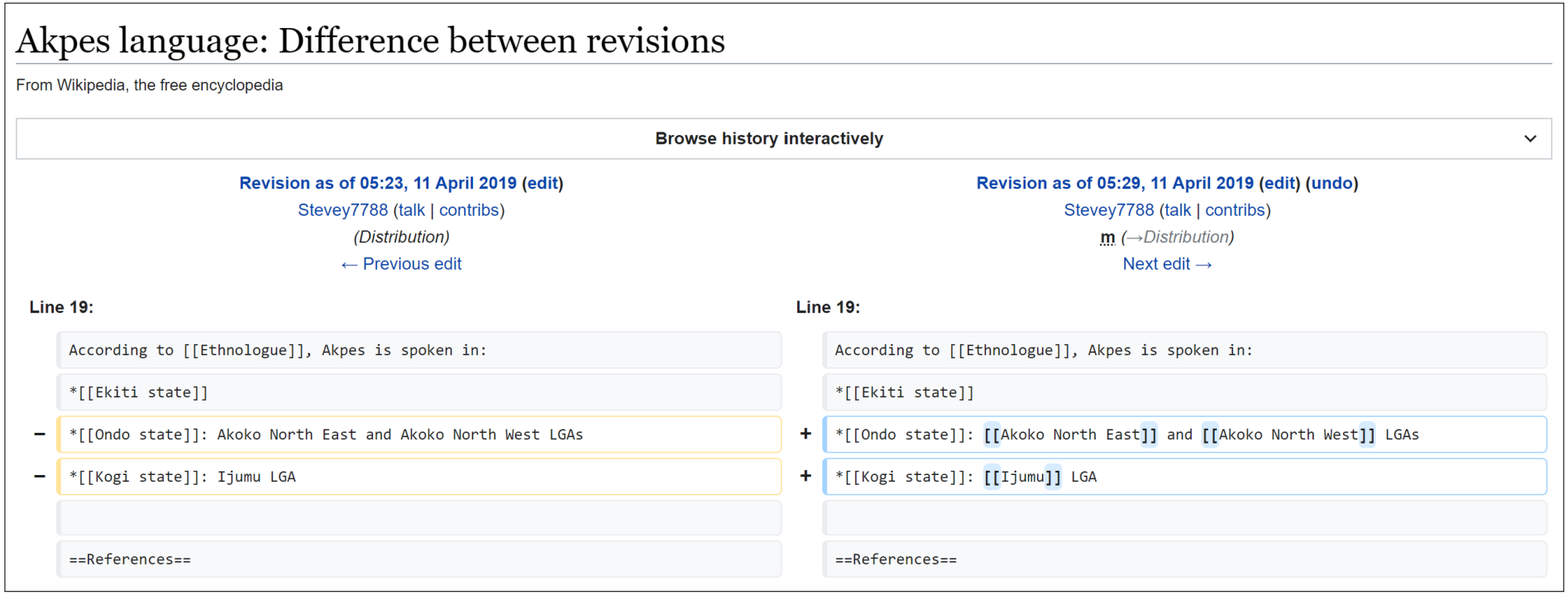


Figure 7.27: Warning about accessing an older version



Akpes language: Difference between revisions

From Wikipedia, the free encyclopedia

Browse history interactively

	Revision as of 05:23, 11 April 2019 (edit) Stevey7788 (talk | contribs) *(Distribution)* ← Previous edit	**Revision as of 05:29, 11 April 2019 (edit) (undo)** Stevey7788 (talk | contribs) m *(→Distribution)* Next edit →
	Line 19:	**Line 19:**
	`According to [[Ethnologue]], Akpes is spoken in:`	`According to [[Ethnologue]], Akpes is spoken in:`
	`*[[Ekiti state]]`	`*[[Ekiti state]]`
− / +	`*[[Ondo state]]: Akoko North East and Akoko North West LGAs`	`*[[Ondo state]]: [[Akoko North East]] and [[Akoko North West]] LGAs`
− / +	`*[[Kogi state]]: Ijumu LGA`	`*[[Kogi state]]: [[Ijumu]] LGA`
	`==References==`	`==References==`

Figure 7.28: Version History Mockup

7.5 Shortcomings of the format and adjustments

In the beginning of Section 7.3, the reasons for implementing a GET were listed (copied below). As mentioned in that section, these reasons can also be considered when evaluating the GET itself and the grammar format.

1. Exemplify a GET and its functionalities
    (a) Test whether the functionalities can be implemented
    (b) Test whether the functionalities improve the reading experience
2. Test the proposed grammar format
    (a) Test the interaction between format and application
    (b) Test whether all content and the represented relations can be extracted from the format
    (c) Find out possible shortcomings of the format

As shown in Section 7.3, many functionalities were implemented in the Abesabesi GET (Reason 1a). The feasibility of the other features has been discussed in Section 7.4. While many of the proposed features should be feasible, the challenges arising when implementing these features have been discussed.

As mentioned before, the reading experience (Reason 1b) will not be evaluated in a quantitative setting. As I have written the grammatical description myself, I do not have a representative opinion on whether the reading experience has improved in comparison to a paper grammar. What can be noted, however, is that a GET offers more navigation tools – especially the search functionalities, access to all databases, and utilization of different layers. These features, together with unimplemented ones such as the TOC alternatives, offer the reader a greater variety of ways to explore the grammar. However, the reading experience might have also been worsened in other ways that would need to be tested on a larger scale.

As the development of the GET and the creation of the format for DRGs happened simultaneously, most shortcomings have already been adjusted. TEI-Grammar, as presented in Section 6.5, has developed to its current form after multiple shortcomings have been adjusted. For instance, every time the interaction between the application and the format did not function (Reason 2a) or specific elements or relations betweens different elements could not be appropriately extracted (Reason 2b), the format was adjusted. This has particularly occurred when elements were not sufficiently connected. Links between different elements could only be established if they were appropriately encoded in the format. The following paragraphs discuss various shortcomings that have been detected while testing the interaction between the GET and the grammar format (Reason 2c).

All parts of the GET that have not been fully implemented might entail undetected shortcomings in respective parts of the format. Finishing the implementation of the GET

would thus improve the quality of TEI-Grammar. This concerns the improvement of the full text search and the implementation of the advanced search in particular. The advanced search harnesses the advantages of semantic markup. In order to test whether the additional encoding of linguistic units is worth the effort, the advanced search functionalities need to be implemented. Inaccurate or false search results could reflect inappropriate and inefficient encoding of linguistic units.

Moreover, the Abesabesi GET still uses an older form of IGT encoding due to time constraints. TEI-IGT thus still needs to be tested in interaction with the application. This task is rather important because it needs to be tested whether the content of TEI-IGT can be read and displayed in an appropriate way. Especially the realization of encoded layout choices might entail potential problems as it is a comparatively complex undertaking.

The processing of the encoded sketch grammar has mostly been hindered by faulty encoding. The XML editor validates TEI documents, but as TEI-Grammar is only a list of recommendations and no schema exists, faulty encoding could not be detected by the editor. However, the format is not meant to be directly used by grammar authors. In an ideal case, a GWT could guide the author and hide the grammar format behind an editor. It would then guarantee a format-compliant encoding to ensure stable processing of the grammatical description by the GET. This could automatize parts of the encoding such as the numbering of grammar entries and other numbered elements. As the numbering is time-consuming and error-prone, I used XML transformations to automatically number the grammar entries.

The discussion about the index in Section 7.3.4.2 has showed that the fusion of index and terminology database works well for this project but should not generally be recommended in the grammar format. This would imply that there should be two separate documents for the terminology database and the index. While the terminology database could be encoded as demonstrated in Section 6.5.4.3, the index could simply be a nested list of terms that refer to their entries in the terminology database.

Chapter 8

Conclusion, outlook, and recommendations

This thesis has presented the full process of grammar writing from the perspective of a documentary linguist. But instead of choosing the traditional medium of grammar writing – a book – I have taken the digital path in order to produce a digital reference grammar. Undertaking the full process of grammar writing within a language documentation setting involved fieldwork, data processing and analysis, and the encoding of the grammatical description. This project has thus produced a case study that, in particular, considered the needs and existing workflows of grammar authors in a documentation setting.

This thesis was divided into two parts. Part I presented the result of the fieldwork and grammar writing process – a sketch grammar of Abesabesi. Part II, on the other hand, dealt with the process of producing a digital reference grammar. Both parts contain separate findings but can be regarded as a contiguous process towards a particular digital reference grammar as well as a general format for digital reference grammars.

8.1 Part I

The first part of this thesis was concerned with the language Abesabesi and its socio-cultural background. Chapter 2 contained a comprehensive overview of available resources on Abesabesi and its speakers. It presented information on the language and its speakers, its location, classification, socio-linguistic setting, and history. Moreover, it described the methods for collecting, processing, and analyzing Abesabesi data.

Chapter 3 provided a primary grammatical description of Abesabesi that included the topics: phonology, word classes, phrase types, clause types, information structure, and the morphology of nouns, verbs, pronouns, and other minor word classes. The typological sketch in Section 2.1.8 can be taken as a summary of the most salient features of Abesabesi. The Abesabesi Sketch Grammar extends the perspective given by Agoyi (2008), who provided the only comprehensive work on Abesabesi so far. This involves

an alternative account on her analysis of vowel harmony in Ekiromi (Section 3.1.4.4), a more in-depth analysis of topics she investigated peripherally (tonology, TAM markers, the pluractionalis marker, and the gerund forms), and an analysis of additional phenomena on different linguistic levels. The sketch grammar has detected phenomena common to the language family and the region (e.g., nominal prefixes, logophoric pronouns, SVO word order, and gerunds) as well as some rather idiosyncratic features, such as the realis/irrealis pronoun sets, the desiderative mood, and closed syllables. Moreover, this project has produced an openly accessible corpus that can be utilized for various research questions and for creating a more comprehensive grammatical description.

The aim of this work was to produce a grammatical description in the from of a sketch grammar. Using the sketch grammar as a base and the compiled corpus as a source for more evidence, future research could result in a more comprehensive or more in-depth grammatical description that would then result in a better understanding of the language's structure. Several new questions concerning Abesabesi's grammar have opened up for future research. The following list presents some examples where topics of the Abesabesi Sketch Grammar need to be investigated further.

1. Abesabesi exhibits high tones to express grammatical relations or categories, such as the interrogative mood, possession, or affirmative polarity. Further research might distinguish more functions of the high tone and clarify constraints or rules.

2. This sketch grammar has differentiated several TAM categories. Their exact semantics, especially in combination with other TAM markers, need to be analyzed.

3. The sketch grammar identified a number of marginal phenomena that occur only infrequently, such as the pluractionalis suffix and the relativized focus construction. Their exact semantics needs to be investigated with a larger corpus.

4. Certain information structure constructions were identified and analyzed on a sentence level. Further research could investigate these constructions and other phenomena on a discourse level.

This work has set a base for the description of many previously undescribed phenomena of Abesabesi. Abesabesi's genetic affiliation has only been discussed based on existing literature. The new data and findings of this project, however, might also enable a more informed perspective on Abesabesi's genetic affiliation.

8.2 Part II

Chapter 4 introduced the field of grammaticography and extended it to include grammars in a digital medium. Chapter 5 then dealt with a conceptualization of how a digital reference grammar (DRG) could function. It laid out the three essential components of a DRG – a grammar writing tool (GWT), grammar documents, and a grammar exploration

tool (GET). The chapter compared traits of a possible DRG with those of existing paper grammars. Having listed what is valued in grammars, it could be differentiated what each grammar medium could offer to follow the respective value. Chapter 6 put forward a format for DRGs – TEI-Grammar. It was tested by encoding the Abesabesi Sketch Grammar and several samples of existing paper grammars. While TEI has been proposed before for the encoding of reference grammars (Drude, 2012; Nordhoff & Hammarström, 2014), this is the first comprehensive approach on a theoretical and practical level. Besides the detailed theoretical discussion on how to encode reference grammars using TEI, this project has also produced practical guidelines that are based on the actual encoding of a digitally born reference grammar and existing paper reference grammar samples.

Furthermore, I stressed the importance of a digital encoding format for grammars and, in particular, the need for a separate TEI module for grammar writing. Although TEI has been proven to suffice for the basic encoding of grammatical descriptions, it could benefit from a separate TEI module in order to guide grammar authors and to result in a tighter and more tailor-made method of encoding. Chapter 7 discussed the development of a GET that reads the format and displays it in a reader-friendly manner. The development helped to detect shortcomings of the format and demonstrated the feasibility of developing a GET.

TEI-Grammar is a rather flexible encoding guideline that uses existing TEI elements. Therefore, it can be used to encode all sorts of reference grammars using the existing TEI infrastructure. The project has also illustrated which components are necessary to structure a reference grammar. It has also emphasized the importance of an integrated approach that connects the grammatical description with other language resources, such as a dictionary, a text collection, or ontologies. This integrated approach fosters the use of consistent terminology and enables the reader to access other relevant resources quickly and to retrieve information. By using existing TEI elements, linguistic units and interlinear examples have emerged as the most idiosyncratic parts of a reference grammar that might need a more refined structure or adjustments in TEI. Adequate adjustments have been proposed in Section 6.5.6. The most central unit of grammar writing – the grammar entry – might require a separate TEI element to enforce a more rigid structure and to differentiate them from ordinary text devisions. The proposed TEI-Grammar, along with these recommendations, could be used as a base for a separate grammar module in TEI.

To enable the encoding of interlinear examples in TEI, I have also proposed TEI-IGT. TEI-IGT is a set of guidelines that contain a method for storing IGT that is based on the logics of Xigt (Goodman et al., 2015). It not only allows for the storage of IGT but also fulfills additional needs related to grammar writing, such as storing layout choices and the grouping of texts. This way, the format can also be used to store interlinear examples and a text collection. The guidelines can further be used beyond grammar writing whenever IGT needs to be stored.

Future research should test the proposed formats on a larger scale. TEI-Grammar and

TEI-IGT have both only been tested on the Abesabesi Sketch Grammar and a small amount of samples from existing grammars. Testing on a larger scale could verify whether the formats can be used for all kinds of reference grammars without restricting the authors too much. This would require a wider range of grammars and a more systematic testing method. Although TEI-Grammar has been designed for digitally born grammars, its usability for retro-digitized paper grammars could also be extended.

Chapter 7 illustrated the development of a GET and investigated what kind of functionalities a GET could have in order to help the reader explore the grammar. It discussed different kinds of navigation tools, access to the databases, the correct display of grammar entries, and the use of hyperlinks to connect different resources.

Not every proposed feature for GETs could be investigated in detail. Some were only discussed briefly because they either do not interact with the format a lot or need to be tested in interaction with a GWT. For instance, this holds true for a version control/history tool and a login facility. Moreover, some features have not been implemented as they were not essential to the discussion (Section 7.4). The full implementation of these features could, however, enrich the application and ensure that they are useful.

Although the GET developed here is only meant as an example and has the primary function of testing the grammar format, it could be the basis for testing the user experience of digital reference grammars. This could ensure that the GET is reader-friendly and accelerates the process of finding information in the grammar.

8.3 Recommendations

DRGs rely on a stable software infrastructure in order for users to author and explore grammatical descriptions. While this project has set a base by providing a format for DRGs, the other two components, GWT and GET, are still missing. Grammar authors cannot be expected to encode their grammars directly into a markup language like XML and develop a GET like it has been done for the Abesabesi Sketch Grammar. The future of digital grammaticography will thus remain uncertain unless there is a larger-scale project to develop a grammar writing infrastructure, as envisioned in Section 5.4.4. Such a project needs to encompass a GWT where all kinds of reference grammars can be authored, server space where grammars can be stored, and a GET where different grammars can be read.

Such a project would require at least one organization to bear the development and maintenance costs. The development would require several developers and staff to maintain the infrastructure. An appropriate organization could either be a commercial company or a research institute. A commercial company could, for example, be a publisher that would let authors use the GWT and let readers pay to explore the different grammars. Although the grammars would not be openly accessible, the publisher could ensure a good quality software and stable maintenance due to the earnings. The Enhanced Electronic Grammars by Mouton De Gruyter showcase a similar model where a publisher processes

and enhances existing grammars and develops a web application to let paying readers explore the grammars (Section 4.6.5). Digital open access publishers, such as Language Science Press, have a different policy and need to rely on grants to maintain the DRG infrastructure. A research organization could also grant open accessibility and would rely on funding. A similar undertaking by a research organization has been the DOBES infrastructure[1]. It has been backed by the Max Planck Institute for Psycholinguistics in Nijmegen. Many functioning software projects have been developed and made openly accessible (ELAN, Arbil, etc.). However, current changes in the internal organization and in the research focus of the institute could possibly also affect the archive infrastructure and the maintenance of the software programs. In order to grant the longevity of a DRG platform, these factors need to be considered.

The institution backing the DRG platform would also be responsible for recommending or enforcing a grammar format (Nordhoff, 2008, p. 319). Depending on the needs of the project, a new standard could be developed or an existing format could be used or extended. Whether a new or an existing format is used, a standard format for DRG would benefit from an institution administering and refining it.

A smaller but possibly more realistic step towards the longevity of DRGs could be a transformation approach using an XSLT document (Section 6.1.1). Such an XSLT document could be easily developed by a researcher and would allow for a transformation of (TEI-Grammar-) encoded grammatical descriptions into a linear XHTML document that can be displayed in a browser. The transformation does not require a custom-made web application to dynamically display the content of the grammar document, but simply relies on an XSLT processor that is part of most XML editors and web browsers. This approach enables a simplified workflow for grammar publishing, as the processor would transform the encoded grammatical description into a static XHTML document only once before putting this XHTML document onto an existing website where it could be accessed by web browsers. The encoded grammatical descriptions could be archived along with their transformed XHTML counterparts to be accessible by researchers in the future.

The disadvantages of such a transformation approach include the fact that authors could not rely on the assistance of a GWT. Authors would need to go through the time-consuming and error-prone encoding process without guidance or automation. This might result in vast differences in how particular grammars are encoded. Moreover, particular features of the GET proposed in this thesis, such as advanced search facilities or version control, would need to be developed separately if required. However, the transformation approach is a feasible solution for standardized DRGs in the future, since it relies on a standardized storage format such as TEI-Grammar and it does not require recurring funding to maintain a centralized DRG infrastructure.

[1]Dokumentation Bedrohter Sprachen (Documentation of Endangered Languages), retrieved 2021-02-05, from `https://dobes.mpi.nl/`

Linguistic glosses

1H	first position high tone
ABSV	active body state verb
AFF	affirmative
ATR	advanced tongue root
C	consonant
CFOC	contrastive focus
CIRCN	circumstantial necessity
COM	comitative
COMP	complementizer
COND	conditional
CONJ	conjunction
CPOR	comparator
DAT	dative
DATO	dative object
DEM	demonstrative
DES	desiderative
DET	determiner
EPN	epistemic necessity
EMPH	emphasis
FAC	factative
FOC	focus
FUT	future

GENH genitive high tone

GRND gerund

H high tone

HAB habitual

HUM human

IDP independent pronoun

IRR irrealis

L low tone

LOC locative

LOCH locative high tone

LOG logophoric

M mid tone

MAINV main verb

N noun

NEG negation

NFUT non-future

NHUM non-human

NMLZ nominalizer

NP noun phrase

NUM numeral

OBJ object

PERS person

PAR partitive

PL plural

PLUR pluractionalis

PN property noun

PO primary object

POSS possessive

POSTV postverb

PP prepositional phrase

PREP preposition

PREV preverb

PRF perfect

PRO pronoun

PROG progressive

PST past

Q interrogative

QH interrogative high tone

RECP reciprocal

REFL reflexive

REL relativizer

RELH relative high tone

SBJ subject

SCM state-changing marker

SEQU sequential

SG singular

SIP sentence-initial particle

SO secondary object

SUP superlative

SVC serial verb construction

TAM tense, aspect, and mood

TBU tone-bearing unit

V vowel

VH vowel harmony

VP verb phrase

Glossary

AIKA Auga, Ishe, Kakumo, and Ayanran: *alternative name for the language Ukaan (ISO 639-3: kcf)*

API Application Programming Interface: *an interface to allow for interaction between different applications*

BLT Basic Linguistic Theory: *a theoretical framework and an ontology of terms that are often used in descriptive linguistics and typology*

CLDF Cross-Linguistic Data Formats: *a set of data formats used for the CLLD databases*

CLLD Cross-Linguistic Linked Data: *a project that aims at the cross linking of linguistic data for language comparison. It manages several databases*

CMS Content Management System: *a type of software that enables users to create and modify digital (web) content without knowledge of programming or markup languages*

CSV Comma Separated Values: *a technical standard that is used to store tabular data. It utilizes the comma to separate values and employs the file extension* `.csv`

DRG Digital Reference Grammar: *a reference grammar stored on a digital medium*

DTD Document Type Definition: *an XML schema language*

EGIDS Extended Graded Intergenerational Disruption Scale: *a scale to assess language vitality developed by M. P. Lewis and Simons (2010)*

ELAR Endangered Languages Archive: *A digital archive that stores deposits of endangered languages. Most of these deposits are funded by ELDP*

ELDP Endangered Languages Documentation Programme: *an organization funded by the Arcadia Fund that provides grants and training for the documentation and preservation of languages*

FLEx Fieldwork Language Explorer: *an application developed by SIL that facilitates the compilation of dictionaries and the glossing of texts*

FOMP Form-Meaning Pair: *a structural unit of grammaticography proposed by Nordhoff (2012b). It is based on the idea that grammatical phenomena can be described as connections between forms and meanings*

GET Grammar Exploration Tool: *a term proposed in this thesis for any tool that enables readers to explore a DRG*

GIDS Graded Intergenerational Disruption Scale: *a scale to assess language vitality developed by Fishman (1991)*

GWT Grammar Writing Tool: *a term proposed in this thesis for any tool that enables authors to produce digitally encoded grammatical descriptions*

HTML Hypertext Markup Language: *a markup language to structure digital documents. It is integral to the structure of websites*

IGT Interlinear Glossed Text: *a method of presenting texts with aligned glosses and a translation into another language*

IPA International Phonetic Alphabet: *an alphabet used to represent speech sounds of the world's languages. It is used for the phonetic or phonemic transcription of languages*

KWIC Key Word In Context: *a method of presenting all occurrences of a keyword along with their immediate contexts*

LGA Local Government Area: *an administrative division of Nigeria below the state level*

NPC Nigerian Population Census: *recurring reports on population figures in Nigeria*

ODD One Document Does it all: *a standard within the TEI framework that is used to generate documentation or schema files*

PDF Portable Document Format: *a software-independent standard for presenting digital text documents*

RDF Resource Description Framework: *a set of standards for the description of resources and their relationships among each other. It is an integral part of the semantic web*

RDFa Resource Description Framework in Attributes: *a standard to use RDF statements in attributes of markup languages such as HTML or XML*

RelaxNG Regular Language for XML Next Generation: *an XML schema language*

RNC Royal Niger Company: *a company charted by the British government that was instrumental in the colonization of Nigeria*

SIL Summer Institute of Linguistics: *an organization that is concerned with the documentation of lesser known languages. It also develops software that aims at facilitating language documentation*

TEI Text Encoding Initiative: *a committee and a technical standard dedicated to the digital encoding of texts*

TEI-Grammar Text Encoding Initiative Reference Grammar Format: *guidelines proposed in this thesis for encoding reference grammars using TEI*

TEI-IGT Text Encoding Initiative Interlinear Glossed Text Format: *guidelines proposed in this thesis for encoding IGT using TEI*

TOC Table of Contents: *a navigation tool in documents that consists of a list of section titles and their corresponding page numbers*

URI Uniform Resource Identifier: *a unique sequence of characters or numbers to identify resources*

URL Uniform Resource Locator: *a special type of URI that localizes a web resource*

W3C World Wide Web Consortium: *an organization dedicated to the development and recommendation of web standards*

WYSIWYG What You See Is What You Get: *a working mode of text editors. WYSIWYG editors allow writers to edit a document while seeing how it would look as a printed copy. Writers are thus not confronted with markup languages*

XHTML Extensible Hypertext Markup Language: *an XML based markup language that mirrors HTML. It is typically used as a transformation product from XML documents*

Xigt Extensible interlinear glossed text: *a standard to store IGT data proposed by (Goodman et al., 2015)*

XML Extensible Markup Language: *a widely-used markup language to store and model textual data in a hierarchical structure*

XQuery XML Query: *a programming language that facilitates querying through XML databases*

XSL Extensible Stylesheet Language: *a series of XML technologies that facilitate the transformation of XML documents*

XSLT Extensible Stylesheet Language Transformations: *a language to transform XML documents into other XML or similar documents such as XHTML*

References

Abimboye, D. (2008). *The History of Ikaramu*. Lambe, Nigeria: Bode Ogbon Communications Services.

Agoyi, T. O. (1998). *Ìsọ̀rì Ọ̀rọ̀ Nínú Èkìròmì* (Unpublished master's thesis). University of Ilorin.

Agoyi, T. O. (2001). The category of number and the genetic classification of Èkìròmì. *Inquiry in African Languages and Literatures*, *4*, 64–80.

Agoyi, T. O. (2008). *The phonology of Àbèsàbèsì vowel harmony* (Unpublished PhD thesis). University of Ilorin.

Agoyi, T. O. (2010). ATR + Roundedness harmony in Àbèsàbèsì: the case study of Akpes. *Akungba Journal of Linguistics and Literature*, *August*, 85–102.

Agoyi, T. O. (2012a). Àbèsàbèsì Orthography: A Proposal. *IOSR Journal Of Humanities And Social Science*, *6*(2), 41–45.

Agoyi, T. O. (2012b). ATR + Round and Low Vowel Harmony in Àbèsàbèsì: A Case Study of Ọ̀ṣùgù. *International Journal of Business and Social Science*, *3*(23), 243–256.

Agoyi, T. O. (2012c). Morphology of Àbèsàbèsì Numerals: A Case Study of Èkìròmì. *International Journal of Business and Social Science*, *3*(20), 231–240.

Agoyi, T. O. (2013a). Classification of Abesabesi. *International Journal of Language and Literatures*, *1*(1), 37–51.

Agoyi, T. O. (2013b). Language, Invasion and Insecurity: The History of Àbèsàbèsì. *American Journal of Social Issues and Humanities*, *3*(2), 68–77.

Agoyi, T. O. (2013c). Ǹń Kpasi Àbèsàbèsì: Mo Gbọ́ Àbèsàbèsì: I Understand Àbèsàbèsì. *International Journal of Humanities and Social Science Invention*, *2*(2), 42–52.

Agoyi, T. O. (2014). Àbèsàbèsì Language Documentation and Maintenance. *The International Journal Of Engineering And Science*, *3*(7), 1–7.

Agoyi, T. O. (2015a). Àbèsàbèsì Syllable and Syllabification: A Generative Phonological Approach. *Ago-Iwoye Journal of Languages and Literary Studies*, *6*, 246–270.

Agoyi, T. O. (2015b). Achieving Millennium Development Goals in Africa

through Indigenous Language-Public Health Policy in Africa: A case Study of Àbèsàbèsì speaking community. *International Journal of Media, Journalism and Mass Communications (IJMJMC)*, *1*(2), 26–35.

Allison, I. O. (2017). Vowel Deletion and Insertion in Úwù. *Journal of Language and Education*, *3*(2), 19–29. doi: https://doi.org/10.17323/2411-7390-2017-3-2-19-29

Ameka, F. K. (2006). Real descriptions: Reflections on native speaker and non-native speaker descriptions of a language. In F. K. Ameka, A. Dench, & N. Evans (Eds.), *Catching language: The standing challenge of grammar writing* (pp. 69–112). Berlin: Mouton de Gruyter.

Ameka, F. K., Dench, A., & Evans, N. (Eds.). (2006). *Catching language: The standing challenge of grammar writing*. Berlin: Mouton de Gruyter. doi: https://doi.org/10.1515/9783110197693

Atoyebi, J. D. (2010). *A Reference Grammar of Oko. A West Benue-Congo Language of North-Central Nigeria*. Köln: Rüdiger Köppe.

Ayoola, O. J. (1986). *Aspects of the Phonology of Daja* (Unpublished bachelor's thesis). University of Ilorin.

Becerra-Bonache, L., Bel-Enguix, G., Jiménez-López, M. D., & Martín-Vide, C. (2015). Mathematical Foundations: Formal Grammars and Languages. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics 2nd edition*. Oxford: Oxford University Press. doi: https://doi.org/10.1093/oxfordhb/9780199573691.013.021

Bender, E. M., Ghodke, S., Baldwin, T., & Dridan, R. (2012). From Database to Treebank: On Enhancing Hypertext Grammars with Grammar Engineering and Treebank Search. In *Electronic Grammaticography* (Vol. 4, pp. 179–206). Retrieved 2021-03-24, from `http://hdl.handle.net/10125/4535`

Bird, S., & Simons, G. (2003). Seven Dimensions of Portability for Language Documentation and Description. *Language*, *79*(3), 557–582. doi: https://doi.org/10.1353/lan.2003.0149

Black, C. A., & Black, H. A. (2012). Grammars for the people, by the people, made easier using PAWS and XLingPaper. In *Electronic Grammaticography* (Vol. 4, pp. 103–128). Retrieved 2021-03-24, from `http://hdl.handle.net/10125/4532`

Black, H. A. (2009). Writing Linguistic Papers in the Third Wave. *SIL Forum for Language Fieldwork*, *004*.

Bow, C., Hughes, B., & Bird, S. (2003). Towards a General Model for Interlinear Text. *Preceedings of the E-MELD Workshop 2003: Digitizing & Annotating Texts and Field Recordings, July, 11-13*. Retrieved 2021-03-24, from `http://emeld.org/workshop/2003/bowbadenbird-paper.html`

Boyd, G., Asiabuka, C. C., Medupin, A., & Osunsanya, A. B. (2009). *Mid-term Assessment of the Millennium Villages Project in Nigeria at Ikaram/Ibaram in Ondo State and at Pampaida in Kaduna State.* Retrieved 2020-03-12, from `https://erc.undp.org/evaluation/evaluations/detail/4149`

Chen, W. (2020). *A Grammar of Southern Min: The Hui'an Dialect.* Berlin/Boston: De Gruyter Mouton.

Chiarcos, C., Hellmann, S., & Nordhoff, S. (2011). Towards a linguistic linked open data cloud: The open linguistics working group. *TAL Traitement Automatique des Langues*, *52*(3), 245–275. doi: https://doi.org/10.1007/978-3-642-31782-8

Chiarcos, C., & Sukhareva, M. (2015). OLiA - Ontologies of Linguistic Annotation. *Semantic Web*, *6*(4), 379–386. doi: https://doi.org/10.3233/SW-140167

Childs, G. T. (2011). *A Grammar of Mani.* Berlin/Boston: Walter de Gruyter.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chovwen, A., Orebiyi, O., Abdou-Salam, S., Afere, T., & Afolayan, E. (2009). Achieving the Millennium Development Goals: An Assessment of Water and Sanitation Intervention of the Ikaram Millennium Village, Nigeria. *Researcher*, *1*(2), 6–13.

Comrie, B. (1976). *Aspect. An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: University Press.

Comrie, B., Croft, W., Lehmann, C., & Zaefferer, D. (1993). A framework for descriptive grammars. In A. Crochetière, J.-C. Boulanger, & C. Ouellon (Eds.), *Proceedings of the XVth International Congress of Linguists (1992)* (pp. 159–170). Sainte-Foy: Les presses de l'université Laval.

Comrie, B., Haspelmath, M., & Bickel, B. (1982). The Leipzig Glossing Rules. *Folia Linguistica*, 1–10.

Comrie, B., & Smith, N. (1977). Lingua Descriptive Studies: Questionnaire. *Lingua*, *42*(1), 1–72. doi: https://doi.org/10.1016/0024-3841(77)90063-8

Cristofaro, S. (2006). The organization of reference grammars: a typologist user's point of view. In F. K. Ameka, A. Dench, & N. Evans (Eds.), *Catching language: The standing challenge of grammar writing* (pp. 137–170). Berlin: Mouton de Gruyter.

Cristofaro, S. (2009). Grammatical Categories and Relations: Universality vs. Language-Specificity and Construction-Specificity. *Language and Linguistics Compass*, *3*(1), 441–479.

Croft, W. (2001). *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Crozier, D. H., & Blench, R. (1992). *An index of Nigerian languages*. Dallas, Texas: Summer Institute of Linguistics.

Crystal, D. (2000). *Language Death*. Cambridge: Cambridge University Press.

Dahl, O. (1985). *Tense and aspect systems*. Oxford: Blackwell.

Daramola, S. (1984). *Aspects of the phonology of Akunnu* (Unpublished bachelor's thesis). University of Ilorin.

de Jong, F. (2007). *A grammar of modern Uyghur*. Utrecht: Houtsma.

Dixon, R. M. W. (1994). *Ergativity*. Cambridge: Cambridge University Press.

Dixon, R. M. W. (2009). *Basic Linguistic Theory, Vol. 1. Methodology*. Oxford: Oxford University Press.

Dixon, R. M. W. (2011). *Basic Linguistic Theory, Vol. 2. Grammatical topics*. Oxford: Oxford University Press.

Dixon, R. M. W. (2012). *Basic Linguistic Theory, Vol. 3. Further grammatical topics*. Oxford: Oxford University Press.

Dobrin, L. M. (n.d.). *The Arapesh Grammar and Digital Language Archive.* Retrieved 2020-09-18, from `http://www.arapesh.org`

Drude, S. (2003). Advanced Glossing - a language documentation format and its implementation with Shoebox. In *Proceedings of the 2002 International Conference on Language Resources and Evaluation (LREC 2002)*. Paris: ELRA. Retrieved 2021-03-24, from `http://hdl.handle.net/11858/00-001M-0000-000F-8A74-E`

Drude, S. (2012). Digital Grammars: Integrating the Wiki/CMS approach with Language Archiving Technology and TEI. In S. Nordhoff (Ed.), *Electronic Grammaticography* (pp. 160–178). Honolulu: University of Hawai'i Press.

Dryer, M. S. (1986). Primary objects, secondary objects and antidative. *Language*, *62*(4), 808–845.

Dryer, M. S. (2006). Descriptive theories, explanatory theories, and basic linguistic theory. In F. K. Ameka, A. Dench, & N. Evans (Eds.), *Catching language: The standing challenge of grammar writing* (pp. 207–234). Berlin: Mouton de Gruyter. doi: https://doi.org/10.1515/9783110197693.207

Elugbe, B. O. (1984). Morphology of the gerund in Degema and its reconstruction in Proto-Edoid. *Studies in African Linguistics*, *15*(1), 77–89.

Elugbe, B. O. (1989). *Comparative Edoid: phonology and lexicon*. Port Harcourt: University of Port Harcourt Press.

Elugbe, B. O. (2001). The classification of Akpes and AIKA. In *31st Colloquium of African languages and linguistics, Leiden, Holland, August 27-29, 2001*.

Elugbe, B. O. (2012). Comparative Akedoid and West Benue-Congo. In *Towards Proto-Niger-Congo: Comparison and Reconstruction, 18-21 September 2012.* Paris: LLACAN.

Evans, N., & Dench, A. (2006). Introduction: Catching language. In F. K. Ameka, A. Dench, & N. Evans (Eds.), *Catching language: The standing challenge of*

grammar writing (pp. 1–40). Berlin: Mouton de Gruyter.

EXIST-DB. (2012). *eXist-db Open Source Native XML Database*. Retrieved 2020-10-28, from http://exist-db.org

Fabre, A. (2016). *Gramática de la lengua nivacle (familia mataguayo, chaco paraguayo)*. München: LINCOM Europa.

Fadoro, J. O. (2013). Phonological Variation in Akokoid. *Journal of Education and Practice*, *4*(11), 56–67.

Farrar, S., & Langendoen, T. (2003). A linguistic ontology for the semantic web. *Glot International*, *7*(3), 97–100.

Fishman, J. A. (1991). *Reversing language shift: theoretical and empirical foundations of assistance to threatened languages*. Clevedon: Multilingual Matters.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., … Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, *5*(1). doi: https://doi.org/10.1038/sdata.2018.205

Forker, D. (2020). *A grammar of Sanzhi Dargwa*. Berlin: Language Science Press.

Good, J. (2004). The Descriptive Grammar as a (Meta)Database. In *EMELD Language Digitization Project Conference 2004*. Retrieved 2021-03-24, from http://emeld.org/workshop/2004/jcgood-paper.html

Good, J. (2012). Deconstructing descriptive grammars. In S. Nordhoff (Ed.), *Electronic Grammaticography* (pp. 2–32). Honolulu: University of Hawai'i Press.

Goodman, M. W., Crowgey, J., Xia, F., & Bender, E. M. (2015). Xigt: extensible interlinear glossed text for natural language processing. *Language Resources and Evaluation*, *49*(2), 455–485. doi: https://doi.org/10.1007/s10579-014-9276-1

Grinevald, C. (2003). Speakers and Documentation of Endangered Languages. *Language Documentation and Description*, *1*, 52–72.

Güldemann, T. (2018). Historical linguistics and genealogical language classification in Africa. In T. Güldemann (Ed.), *The Languages and Linguistics of Africa* (pp. 58–444). Berlin/Boston: De Gruyter. doi: https://doi.org/10.1515/9783110421668-002

Hansford, K., Bendor-Samuel, J., & Stanford, R. (1976). *An Index of Nigerian Languages*. Accra: Summer Institute of Linguistics.

Haspelmath, M. (1993). *A grammar of Lezgian*. Berlin: Mouton.

Haspelmath, M. (2007). Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, *11*(1), 119–132. doi: https://doi.org/10.1515/LINGTY.2007.011

Heath, J. (1980). *Nunggubuyu Myths and Ethnographic Texts*. Canberra: Australian

Institute of Aboriginal Studies.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, *36*(1998), 161–195.

Hoffmann, C. (1976). The languages of Nigeria by language family. In K. Hansford, J. Bendor-Samuel, & R. Stanford (Eds.), *An Index of Nigerian Languages* (pp. 169–190). Accra: Summer Institute of Linguistics.

Hughes, B., Bow, C., & Bird, S. (2004). Functional requirements for an interlinear text editor. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.* Retrieved 2021-03-24, from `http://lrec-conf.org/proceedings/lrec2004/pdf/273.pdf`

Ibrahim-Arirabiyi, F. (1989). *A comparative reconstruction of Akpes lects, Akoko North, Ondo State* (Unpublished master's thesis). University of Port Harcourt.

Junker, M.-O. (2000). *The Interactive East Cree Reference Grammar.* Retrieved 2020-09-18, from `https://www.eastcree.org/cree/en/grammar/`

Kager, R. (1999). *Optimality Theory*. Cambridge: Cambridge University Press.

Kay, P., & Fillmore, C. J. (1999). Grammatical Constructions and Linguistic Generalizations: The What's X Doing Y? Construction. *Language*, *75*(1), 1–33. doi: https://doi.org/10.2307/417472

Knuth, D. (1984). *The TeXbook*. Reading: Addison-Wesley.

Krauss, M. (1992). The world's languages in crisis. *Language*, *68*(1).

Kruspe, N. (2004). *A grammar of Semelai*. Cambridge: Cambridge University Press.

Langacker, R. W. (2008). *Cognitive Grammar: A basic introduction*. Oxford: Oxford University Press.

Leech, G., & Svartvik, J. (1975). *A communicative grammar of English*. London: Longman.

Lehmann, C. (1982). Directions for Interlinear Morphemic Translations. *Folia Linguistica (Acta Societatis Linguisticae Europaeae)*, *XVI*, 199–224.

Lehmann, C. (2004a). Funktionale Grammatikographie. In W. Premper (Ed.), *Dimensionen und Kontinua. Beiträge zu Hansjakob Seilers Universalienforschung* (Diversitas ed., pp. 147–165). Bochum: N. Brockmeyer.

Lehmann, C. (2004b). Interlinear morphemic glossing. In G. Booij, C. Lehmann, J. Mugdan, & S. Skopeteas (Eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2. Halbband* (pp. 1834–1857). Berlin: Walter de Gruyter. Retrieved 2021-03-24, from `http://www.christianlehmann.eu/publ/lehmann_img.pdf`

Lehmann, C., Chiarcos, C., & Hellmann, S. (2017). Language, Data, and Knowledge. In *Proceedings of Language, data, and knowledge. First International*

Conference, LDK 2017, Galway, Ireland, June 19-20, 2017 (Vol. 10318, pp. 42–57). Berlin/Heidelberg: Springer. doi: https://doi.org/10.1007/978-3-319-59888-8

Lehmann, C., & Maslova, E. (2004). Grammticography. In G. Booij, C. Lehmann, J. Mugdan, & S. Skopeteas (Eds.), *Morphologie. ein internationales handbuch zur flexion und wortbildung. 2. halbband* (pp. 1857–1882). Berlin/New York: Walter de Gruyter.

Leitner, G., & Graustein, G. (Eds.). (1989). *Reference Grammars and Modern Linguistic Theory*. Berlin/Boston: De Gruyter.

Lewis, M. P., & Simons, G. F. (2010). Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique*, *55*(2), 103–120.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (2018). *Ethnologue: Languages of the World*. Dallas, Texas: SIL International. Retrieved 2021-03-24, from http://www.ethnologue.com

Lewis, W. D., & Xia, F. (2010). Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*, *25*(3), 303–319. doi: https://doi.org/10.1093/llc/fqq006

Matter, F. (2019). Using CLDF and CLLD for language documentation and description. In *UPPLADOC*. Uppsala. Retrieved 2021-02-17, from https://florianmatter.gitlab.io/pdfs/uppladoc_gramr.pdf

Matter, F. (2020). Combining language description and documentation. In *Webinar*. Lund: Center for Languages and Literature. Retrieved 2021-02-17, from https://florianmatter.gitlab.io/pdfs/gramr_lund_slides.pdf

Maxwell, M. (2011). Electronic Grammars : Taking advantage of the possibilities. In *2nd international conference on language documentation and conservation (icldc)*. Honolulu. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/5189/2/Maxwell-c4.pdf

Maxwell, M. (2012). Electronic Grammars and Reproducible Research. In S. Nordhoff (Ed.), *Electronic Grammaticography*. Honolulu: University of Hawai'i Press.

McGill, S. (2007). Documenting grammatical tone using Toolbox: an evaluation of Buseman's interlinearisation technique. In P. K. Austin, O. Bond, & D. Nathan (Eds.), *Proceedings of Conference on Language Documentation and Linguistic Theory* (pp. 177–186). London: SOAS.

Mosel, U. (1977). *Tolai Texts. Kivung 10/1-2*. Port Moresby: Linguistic Society of Papua New Guinea.

Mosel, U. (1987). *Inhalt und Aufbau deskriptiver Grammatiken*. Köln: Seminar für Sprachwissenschaft.

Mosel, U. (2006). Grammaticography: The Art and Craft of Writing Grammars. In F. K. Ameka, A. Dench, & N. Evans (Eds.), *Catching language: The standing challenge of grammar writing* (pp. 41–68). Berlin: Mouton de Gruyter.

Musgrave, S., & Thieberger, N. (2012). Language description and hypertext: Nunggubuyu as a case study. In *Electronic Grammaticography* (Vol. 4, pp. 63–77).

Nakayama, T., & Rice, K. (2014). *The art and practice of grammar writing*. Honolulu: University of Hawai'i Press. Retrieved 2021-03-24, from `http://nflrc.hawaii.edu/ldc/?p=715`

Newman, P. (1990). *Nominal and Verbal Plurality in Chadic*. Berlin: Mouton de Gruyter.

Nichols, J. (2011). *Ingush Grammar*. Berkeley: University of California Press.

Noonan, M. (2005). Grammar writing for a grammar-reading audience. *Studies in Language*, *30*(2), 351–365. doi: https://doi.org/10.1075/sl.30.2.08noo

Nordhoff, S. (2008). Electronic reference grammars for typology: Challenges and solutions. *Language Documentation & Conservation*, *2*(2), 296–324.

Nordhoff, S. (2009). *A grammar of Upcountry Sri Lanka Malay* (PhD thesis, Universiteit van Amsterdam). Retrieved 2021-03-24, from `https://www.lotpublications.nl/Documents/226_fulltext.pdf`

Nordhoff, S. (Ed.). (2012a). *Electronic Grammaticography. Language Documentation and Conservation Special Publication No. 4*. Honolulu: University of Hawai'i Press.

Nordhoff, S. (2012b). The grammatical description as a collection of form-meaning-pairs. In S. Nordhoff (Ed.), *Electronic Grammaticography* (pp. 33–62). Honolulu: University of Hawai'i Press.

Nordhoff, S., & Hammarström, H. (2014). Archiving grammatical descriptions. *Language Documentation and Description*, *12*, 164–186.

Ohiri-Aniche, C. (1999). Language diversification in the Akoko area of Western Nigeria. In R. Blench & M. Spriggs (Eds.), *Archaeology and Language IV. Language change and cultural transformation* (pp. 79–94). London, New York: Routledge.

Olúmúyìwá, T., & Oshòdì, B. (2012). On the linguistic situation in Àkókó. *California Linguistic Notes*, *XXXVII*(1), 1–8.

Oyebade, F. O. (2015). Morphological Evidence for the Computation of Numbers in Edoid, Akokoid, Akpes, Nupoid and Yoruboid. In O. O. Orie, J. F. Ilọri, & L. C. Yuka (Eds.), *Current Research in African Linguistics: Papers in Honor of Oladele Awobuluyi* (pp. 195–218). Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Oyetade, S. O. (2004). Language Endangerment in Nigeria: Perspectives with

the Akpes Cluster of Akoko Languages. In A. Akinlabi & O. Adesola (Eds.), *Proceedings of the 4th World Congress of African Linguistics New Brunswick 2003*. Cologne: Rüdiger Köppe.

Palmer, A., & Erk, K. (2007). IGT-XML: An XML format for interlinearized glossed texts. In *ACL 2007: The LAW - Proceedings of The Linguistic Annotation Workshop* (pp. 176–183).

Pawley, A. (2014). Grammar writing from a dissertation advisor's perspective. In T. Nakayama & K. Rice (Eds.), *The art and practice of grammar writing* (Vol. 8, pp. 7–23). Honolulu: University of Hawai'i Press.

Payne, T. E. (1997). *Describing morphosyntax: a guide for field linguists*. Cambridge: Cambridge University Press.

Payne, T. E. (2005a). A grammar as a communicative act or What does a grammatical description really describe? *Studies in Language*, *30*(2), 367–383. doi: https://doi.org/10.1075/sl.30.2.09pay

Payne, T. E. (2005b). Introduction: Perspectives on Grammar Writing. *Studies in Language*, *30*(2), 235–243. doi: https://doi.org/10.1177/0002764205283560

Payne, T. E. (2014). Toward a balanced grammatical description. In T. Nakayama & K. Rice (Eds.), *The art and practice of grammar writing* (Vol. 8, pp. 91–108). Honolulu: University of Hawai'i Press.

Payne, T. E., & Weber, D. J. (Eds.). (2007). *Perspectives on grammar writing*. Amsterdam: John Benjamins Publishing Company.

Peterson, J. (2005). *AVG 2.0. Cross-linguistic Reference Grammar. Final Report* (Vol. 008). Universität München: Centrum für Informations- und Sprachverarbeitung. Retrieved 2021-03-24, from `http://www.cis.uni-muenchen.de/download/cis-berichte/02-130.pdf`

Priestley, C. (2020). *Koromu (Kesawai): Grammar and Information Structure of a New Guinea Language*. Berlin/Boston: De Gruyter Mouton.

Raji, B. T. (1986). *Ikaramu phonology* (Unpublished master's thesis). University of Ilorin.

Reiman, D. (2010). Basic oral language documentation. *Language Documentation & Conservation*, *4*, 254–268.

Rice, K. (2005). A typology of good grammars. *Studies in Language*, *30*(2), 385–415. doi: https://doi.org/10.1075/sl.30.2.10ric

Rice, K. (2006). Let the Language Tell Its Story? The Role of Linguistic Theory in Writing Grammars. In F. K. Ameka, A. Dench, & N. Evans (Eds.), *Catching language: The standing challenge of grammar writing* (pp. 235–268). Berlin: Mouton de Gruyter.

Rizzolatti, G., & Craighero, L. (2007). Language and mirror neurons. In

M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press.

Round, E. R., Macklin-cordes, J. L., Ellison, T. M., & Beniamine, S. (2020). Automated Parsing of Interlinear Glossed Text From Page Images of Grammatical Descriptions. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 2878–2883).

Sachnine, M. (2014). *Grammaire du yorùbá standard*. Paris: Karthala.

Salffner, S. (2009). *Tone in the phonology, lexicon and grammar of Ikaan* (PhD thesis, SOAS University of London). Retrieved 2021-03-24, from http://hdl.handle.net/11707/2182

Salffner, S. (2010). *Ikaan and related dialects of Ukaan: an archive of language and cultural material from the Akaan people of Ikakumo (Ondo State, Nigeria)*. Retrieved 2021-03-24, from http://hdl.handle.net/2196/eaacb80f-8877-4cea-8573-63c9f526393b

Salffner, S., & Sands, B. (2012). Why Ukaan is Hard to Classify. In *Towards Proto-Niger-Congo: Comparison and Reconstruction, 18-21 September 2012* (pp. 1–9). Paris: LLACAN.

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt.

Schokkin, D. (2020). *A Grammar of Paluai: The Language of Baluan Island, Papua New Guinea*. Berlin/Boston: De Gruyter Mouton.

Schultze-Berndt, E. (2006). Taking a closer look at function verbs: Lexicon, grammar, or both? In F. K. Ameka, A. Dench, & N. Evans (Eds.), *Catching language: The standing challenge of grammar writing* (pp. 359–391). Berlin: Mouton de Gruyter.

Siegel, E., & Retter, A. (2015). *eXist*. Sebastopol: O'Reilly Media.

Snijders, M. K., Windhouwer, M., Wittenburg, P., & Wright, S. E. (2009). ISOcat: remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies, 4*(4), 261. doi: https://doi.org/10.1504/IJMSO.2009.029230

Suttles, W. P. (2004). *Musqueam reference grammar*. Vancouver: UBC Press.

TEI Consortium. (2020). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.0.0*. Retrieved 2020-08-03, from http://www.tei-c.org/Guidelines/P5/

Thompson, H.-R. (2012). *Bengali*. Amsterdam/Philadelphia: John Benjamins Pub. Co.

UNESCO. (2003). *Language Vitality and Endangerment*. Retrieved 2020-03-06, from http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/Language_vitality_and_endangerment_EN.pdf

von Böhtlingk, O. (Ed.). (1964). *Pâṇini's Grammatik*. Hildesheim: Olms.

W3C. (2017). *XQuery 3.1: An XML Query Language*. Retrieved 2020-10-28, from https://www.w3.org/TR/xquery-31/

Weber, D. J. (2005a). The linguistic example. *Studies in Language*, *30*(2), 445–460. doi: https://doi.org/10.1075/sl.30.2.12web

Weber, D. J. (2005b). Thoughts on growing a grammar. *Studies in Language*, *30*(2), 417–444. doi: https://doi.org/10.1075/sl.30.2.11web

Welmers, W. E. (1974). *African Language Structures*. Berkeley: University of California Press.

Williamson, K. (1989). Benue-Congo overview. In J. Bendor-Samuel (Ed.), *The Niger-Congo languages* (pp. 246–274). Lanham: University Press of America.

Williamson, K., & Blench, R. (2000). Niger-Congo. In B. Heine & D. Nurse (Eds.), *African Languages. An Introduction* (pp. 11–42). Cambridge: Cambridge University Press.

Woodbury, A. (2003). Defining Documentary Linguistics. *Language Documentation and Description*, *1*, 35–51.

Wurm, S. A. (1998). Methods of language maintenance and revival, with selected cases of language endangerment in the world. In K. Matsumura (Ed.), *Studies in Endangered Languages* (pp. 191–211). Hitzzi Syobo.

Yusuf, O. (1990). Yoruba Copula ni. *Journal of West African Languages*, *20*(1), 83–93.

Zaefferer, D. (1998). *Deskriptive Grammatik und allgemeiner Sprachvergleich*. Tübingen: Niemeyer.

Zaefferer, D. (2006). Realizing Humboldt's dream: Cross-linguistic grammatography as data-base creation. In F. K. Ameka, A. Dench, & N. Evans (Eds.), *Catching language: The standing challenge of grammar writing* (pp. 113–135). Berlin: Mouton de Gruyter.

Appendix A

XML/TEI Extracts

Listing A.1: Exemplary XML fragement of a reference grammar's structure (Good, 2004, Section 4)

```
<grammar>
    <ontology id="GOLD" level="general">
        An internal general ontology, or a reference to an
        external general ontology would be placed in this element.
    </ontology>
    <ontology id="MySubcommOnt" level="subcommunity">
        An internal subcommunity ontology, or a reference to an
        external subcommunity ontology would be placed in this
        element.
    </ontology>
    <ontology id="MyLang" level="local">
        An internal local ontology, or a reference to an external
        local ontology would be placed in this element.
    </ontology>
    <ontology id="MiscTerms" level="other">
        It might also be worthwhile to allow for other types of
        ontologies than the three found in the surveyed grammars.
    </ontology>

    <annotation title="Sample annotation" id="annotation_1">
        <ontRef ontologyName="GOLD" ref="some_GOLD_id">
            An annotation can be associated with a reference to an
            ontology. This is a reference to a term from a general
            ontology.
        </ontRef>
        <descProse>
```

```
            Descriptive prose for an annotation would be placed
            here. In addition, there could be inline references to
            a lexical item via an element like the following
            <lexRef ref="some_lexicon_id"/>. There can also be an
            exemplar set using the markup immediately below. The
            descriptive prose could also draw a term from an
            ontology by using an ontology reference as follows
            <ontRef ontologyName="GOLD"
            ref="some_other_GOLD_id"/>.
        </descProse>
        <exSet id="exemplar_set_1">
            <ontRef ontologyName="GOLD" ref="yet_another_GOLD_id">
                Explicit references to ontologies can also be
                placed within example sets.
            </ontRef>
            <ontRef ontologyName="MySubcommOnt"
            ref="some_subcommOnt_id">
                Multiple ontology references are allowed. This is
                a reference to a subcommunity ontology.
            </ontRef>
            <textEx id="some_text_id">
                Content retrieval could be completely automatic or
                could also be specified within the element.
                <ontRef ontologyName="MyLang"
                ref="some_localOnt_id">
                    Ontology references can also be directly
                    related to exemplars. This is a reference to a
                    local ontology.
                </ontRef>
            </textEx>
            <textEx id="some_other_text_id"></textEx>
        </exSet>
        <descProse>
            The exemplars above are textual exemplars. Lexical
            exemplars are also possible, as seen below.
        </descProse>
        <exSet id="exemplar_set_2">
            <lexEx ref="some_lexicon_id"></lexEx>
            <lexEx ref="some_other_lexicon_id"></lexEx>
        </exSet>
```

```
            <annotation title="Sample sub-annotation"
            id="annotation_2">
                <descProse>
                    This is a nested annotation. Here's a reference to
                    the higher-level annotation
                    <crossRef ref="annotation_1"/>, and here's a
                    reference to the textual exemplar set above
                    <crossRef ref="exemplar_set_1"/>.
                </descProse>
            </annotation>

        </annotation>

        <lexicon>
            An internal lexicon, or reference to an external lexicon,
            would be placed in this element.
        </lexicon>

        <texts>
            An internal set of texts, or reference to an external set
            of texts, would be placed in this element.
        </texts>

</grammar>
```

Listing A.2: Encoding of text metadata in TEI-IGT

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
    <!--Metadata for the whole corpus-->
    <teiHeader>...</teiHeader>
    <!--First text-->
    <TEI xmlns="http://www.tei-c.org/ns/1.0">
        <!--Metadata for the first text-->
        <teiHeader>
            <fileDesc>
                <titleStmt>
                    <title>The pregnant tortoise</title>
                </titleStmt>
                ...
                <sourceDesc>
                    <recordingStmt>
                        <recording type="audio">
                            <media url="http://hdl.handle.net/2196/
```

```
                                        00-0000-0000-0013-F5CE-D"
                                        mimeType="audio/wav"/>
                                </recording>
                        </recordingStmt>
                </sourceDesc>
        </fileDesc>
        <profileDesc>
                <particDesc>
                        <person xml:id="AA" sex="M" age="mid">
                                <persName>Adekunle</persName>
                                <birth when="1946-06-01"/>
                                <langKnowledge>
                                <langKnown level="first" tag="ibe">
                                Abesabesi
                                </langKnown>
                                <langKnown level="second" tag="en">
                                English
                                </langKnown>
                                </langKnowledge>
                                <residence>Ikaram-Akoko</residence>
                                <education>BA</education>
                                <occupation>Retired</occupation>
                        </person>
                </particDesc>
                <settingDesc>
                        <setting who="\#AA">
                                <name type="city">Ikaram-Akoko</name>
                                <name type="country">Nigeria</name>
                                <date>2019-02-19</date>
                                <activity>Story-telling</activity>
                        </setting>
                </settingDesc>
                <langUsage>
                        <language ident="ibe" usage="9">Abesabesi</language>
                        <language ident="en-US" usage="90">
                        American English
                        </language>
                </langUsage>
                <textDesc>
                        <channel>face-to-face</channel>
                        <constitution>complete</constitution>
```

```
                                    <derivation type="original"/>
                                    <domain type="private"/>
                                    <factuality type="fiction"/>
                                    <interaction type="none" active="singular"
                                    passive="single"/>
                                    <preparedness type="none"/>
                                    <purpose type="entertain"/>
                            </textDesc>
                    </profileDesc>
            </teiHeader>
            <!--The actual text with annotation-->
            <text>
                    <body>
<!-- IGT instance with corresponding recorded utterance and time frame-->
                            <annotationBlock xml:id="igt01"
                            corresp="texts.xml\#ibe119-00.015"
                            start="16900" end="17980" who="\#AA">
                                    ...
                            </annotationBlock>
                            <annotationBlock xml:id="igt02" ...>
                                    ...
                            </annotationBlock>
                            ...
                    </body>
            </text>
    </TEI>
    <!--other texts-->
    <TEI>
            <teiHeader>
                    ...
            </teiHeader>
            <text>
                    ...
            </text>
    </TEI>
    ...
</teiCorpus>
```

Appendix B

Lists of Nigerian names

The following lists contain names of places and languages in Nigeria. As these names have been written without accent marks and other diacritics throughout this thesis (Section 1.2), they are given here in the Pan-Nigerian Alphabet (mostly in the Yoruba Orthography) in order to indicate the tone and ATR value of vowels. The first column contains the names in the Pan-Nigerian Alphabet. If available, the pronunciation in Abesabesi (taken from the corpus) is given in a third column. Sources of the first column are given in the final column. Diacritics are prone to typing errors in both the sources and the transmission to this list. Therefore, this list is to be used with caution. However, it could serve to identify the entities mentioned within the text and on the maps of this thesis.

Table B.1: Abesabesi communities

Yoruba	Abesabesi	Sources
Àkùnnù	/òkùn/	Agoyi (2013b, p. 75); Olúmúyìwá and Oshòdì (2012, p. 5)
Àsẹ̀	/àsɛ̀/	Agoyi (2013b, p. 75)
Dája	/dádʒa/	Agoyi (2013b, p. 75); Olúmúyìwá and Oshòdì (2012, p. 5)
Ẹ̀ṣùkù	/ɔʃùgù/	Agoyi (2013b, p. 75); Olúmúyìwá and Oshòdì (2012, p. 5)
Gèdègédé	/gèdègédé/	Agoyi (2013b, p. 75); Olúmúyìwá and Oshòdì (2012, p. 5)
Ìbáràmù	/ìbaram̀/	Agoyi (2013b, p. 75); Olúmúyìwá and Oshòdì (2012, p. 5))
Ìkáràmù	/èkiròm/	Agoyi (2013b, p. 75); Olúmúyìwá and Oshòdì (2012, p. 5)
Ìlúdọ̀tun / Ìlọ̀dùn / Àkùnnù Àjọwá	/ìlúdɔ̀tũ/	Agoyi (2014, p. 4); Olúmúyìwá and Oshòdì (2012, p. 5); Agoyi (2014, p. 68)
Ìyànì	/ìjànì/	Agoyi (2013b, p. 75); Olúmúyìwá and Oshòdì (2012, p. 5)

Table B.2: Quarters of Ikaramu

Yoruba	Abesabesi	Sources
Àyétẹ́jú	/ɔʃùgù/	Agoyi (2013b, p. 70)
Ìtabalógun	/ìtùjò/	Agoyi (2001, p. 65)
Ìyọ̀kẹ	/ìjɔ̀kɛ/	Agoyi (2013b, p. 70)
Ìyòtù	/ìjòtù/	Agoyi (2013b, p. 71)
Ọ̀yàgì	/ɔjàgì/	Agoyi (2013b, p. 70)

Table B.3: Abesabesi dialects

Pan-Nigerian Alphabet	Sources
Akpes	Agoyi (2013b, p. 75)
Èkiròmì	Agoyi (2008, Abstract)
Ìluẹ̀nì	Agoyi (2013b, p. 75)
Ọ̀ṣùgù	Agoyi (2008, Abstract)

Table B.4: Places and languages of the Akoko region

Pan-Nigerian Alphabet	Sources
Àbèsàbèsì	Agoyi (2008)
Àfìn	Olúmúyìwá and Oshòdì (2012, p. 5)
Àjọ́wá	Olúmúyìwá and Oshòdì (2012, p. 6)
Àkókó	Olúmúyìwá and Oshòdì (2012, p. 6)
Àkùngbá	Olúmúyìwá and Oshòdì (2012, p. 4)
Arigidi	Olúmúyìwá and Oshòdì (2012, p. 6)
Àúga	Olúmúyìwá and Oshòdì (2012, p. 2)
Àyèré	Allison (2017, p. 19)
Ẹ̀pìnmì	Elugbe (1989, p. 13)
Erúṣú	Olúmúyìwá and Oshòdì (2012, p. 6)
Ìfira	Olúmúyìwá and Oshòdì (2012, p. 4)
Ìgáṣí	Agoyi and Oyebade (p.c.)
Ìkákùmọ	Olúmúyìwá and Oshòdì (2012, p. 2)

Ìkàrẹ́	Olúmúyìwá and Oshòdì (2012, p. 4)
Ikún	Olúmúyìwá and Oshòdì (2012, p. 4)
Imerì	Olúmúyìwá and Oshòdì (2012, p. 4)
Ìpè	Elugbe (1989, p. 13)
Ìpèsì	Olúmúyìwá and Oshòdì (2012, p. 4)
Ìrùn	Olúmúyìwá and Oshòdì (2012, p. 5)
Iṣè	Olúmúyìwá and Oshòdì (2012, p. 2)
Ìṣùà	Elugbe (1989, p. 13)
Ìwárọ̀ Ọ̀kà	Olúmúyìwá and Oshòdì (2012, p. 4)
Ọbà	Olúmúyìwá and Oshòdì (2012, p. 4)
Ọ̀gbàgì	Olúmúyìwá and Oshòdì (2012, p. 5)
Ọ̀jọ̀ Àjọwá	Olúmúyìwá and Oshòdì (2012, p. 6)
Ọ̀kà	Olúmúyìwá and Oshòdì (2012, p. 4)
Òkè-Àgbè	Olúmúyìwá and Oshòdì (2012, p. 6)
Ọ̀kọ	Atoyebi (2010)
Ọ̀sẹ́ Ọbà	Olúmúyìwá and Oshòdì (2012, p. 4)
Ọwọn	Fadoro (2013, p. 56)
Oyín	Agoyi and Oyebade (p.c.)
Ṣósan	Elugbe (1989, p. 13)
Ṣúpárè	Olúmúyìwá and Oshòdì (2012, p. 4)
Ùgbẹ̀	Olúmúyìwá and Oshòdì (2012, p. 4)
Ùrò Àjọwá	Agoyi and Oyebade (p.c.)
Úwù	Allison (2017, p. 19)