ICAME 41 Heidelberg Digital Conference

# Language and Linguistics in a Complex World

## Data, Interdisciplinarity, Transfer, and the Next Generation

The International Computer Archive of Modern and Medieval
English annual conference
Heidelberg University and University of Cologne, Germany
May 20-23, 2020

## EXTENDED BOOK OF ABSTRACTS

Conference Organizers: Beatrix Busse, Ingo Kleiber, Nina Dumrukcic, Franziska Wagner, Ruth Möhlig-Falke, and Kellie Gonçalves.
Edited by Beatrix Busse, Nina Dumrukcic, and Ruth Möhlig-Falke

ICAME41 Heidelberg Digital Conference
Heidelberg University and University of Cologne, Germany

Published by University and City Library (USB) Cologne

Editors:
Beatrix Busse
Nina Dumrukcic
Ruth Möhlig-Falke

Cologne University Publication Server KUPS
Universitäts- und Stadtbibliothek Köln
Universitätsstr. 33
50931 Köln
https://kups.ub.uni-koeln.de/

Conference organized with support from
DFG Deutsche Forschungsgemeinschaft (German Research Foundation)
John Benjamins Publishing Company
University of Cologne

# Preface

ICAME conferences aim to provide the research community with a view of the state-of-the-art in English corpus linguistics and corpus linguistics in general. For ICAME 41 our aim was to go beyond this: we wanted to discuss how we can take corpus linguistics out of its comfort zone and realize its (inter-)disciplinary potential. Given the great societal challenges we are currently experiencing, Harari (2019, p. 266) has claimed that "the heat is on" in teaching and education to find solutions. As a result of, for example, (hyper)globalization, digitization, and developments in artificial intelligence and machine-learning, including big-data research, we are faced with a growing global and local complexity and interdependence of matter, lives, people and things, which also includes language. However, we have not even begun to understand how all of these issues and concepts will interact (humanely, sensibly, peacefully and sustainably) under the new circumstances of rapid change - nor have we yet considered what role linguistics and corpus linguistics may have to play in the solution of global challenges, and how education and teaching will consequently have to be transformed in order to do this.

Language is everywhere, we use language(s) every day, and language and discourse shape our thinking as well as our lives. And yet, the general public has little knowledge of what linguistics and corpus linguistics is and does. Even though (historical) corpus linguistics has produced pioneering research findings on (quantitative and qualitative) linguistic and other semiotic patterns of language use for more than half a century, and has compiled various kinds of groundbreaking synchronic and diachronic language corpora and developed software tools, the impact of corpus linguistics is still fairly low and its social functions have not been sketched, let alone realized. Therefore, we think there is a need to transfer the methods and insights of corpus linguistics and its related sub-branches to society more forcefully than has been attempted in the past. This involves making accessible the excellent digital research on language, as well as the software, tools and methods of analyzing linguistic data big and small, past and present. Corpus linguistics needs to become more interdisciplinary to contribute to solving the grand societal challenges and to emphasize that language is the crucial social and cultural factor in human interaction. At the same time, the corpus linguistic community needs to take on responsibility for educating and training the next generation, with these challenges in mind as they do so.

Hence, the theme we chose for this conference was *Language and Linguistics in a Complex World: Data, Interdisciplinarity, Transfer, and the Next Generation*.

Besides facilitating the exchange and discussion of cutting-edge research, our aims were as follows:

- make participants reflect upon the relevance of their research for the community and society generally, and consider how it can be transferred through, for example, communication, counseling or even patent management;
- make participants think about ways in which we can join forces to find answers to the grand societal challenges, which are always interdisciplinary and can only be addressed in teams. Applying our methodologies and approaches to 'real-world problems' and communicating well the value that corpus linguistics can bring to different areas and domains will also facilitate the renewal of corpus linguistics as a methodology/discipline;
- critically discuss the interplay between artificial intelligence and corpus linguistics and address the growing divide between 'traditional' corpus linguistics and primarily tech-driven emerging fields such as 'data science' and 'natural language processing';
- set up topic-related working groups of senior and junior experts that address these aims and volunteer, for example, to prepare innovative (digital) teaching and learning material and environments for the next generation of students of corpus linguistics and junior researchers.

The process of moving the conference into the digital space and what it takes to organize a digital conference is described in more detail in Busse and Kleiber (2020). Instead of publishing a book of abstract as is the norm for academic conferences, we wanted our participants to describe their work in a compact yet informative way. This is a collection of papers, work-in-progress reports, and other contributions that were part of the ICAME41 digital conference. We would like to sincerely thank Heidelberg University, the University of Cologne, as well as the DFG Deutsche Forschungsgemeinschaft (German Research Foundation) and the John Benjamins Publishing Company for their support.

Beatrix Busse and the ICAME41 team

## *References*

Busse, B., & Kleiber, I. (2020). Realizing an online conference: Organization, management, tools, communication, and co-creation. *International Journal of Corpus Linguistics, 25*(3), 322-346.

**Ajšić, Adnan**

***American University of Sharjah***

**aajsic@aus.edu**

Type of Contribution: Work-in-Progress Report

## Large-scale lexical patterns in discourse as indicators of ideologies

Comparing exploratory factor analysis and topic modeling

*Abstract*

Recent (lexical) approaches to identification of discourses (e.g. Baker et al., 2008; Partington, 2010), and language ideologies in particular (e.g. Subtirelu, 2013, 2015; Vessey, 2017), focus on the application of quantitative corpus-linguistic techniques to large data sets as a way to ensure more objective sampling methods and replicability of analytical procedures, as well as to minimize researcher inference. In addition to studies applying exploratory factor analysis (Ajšić, 2015, 2016, 2021; Berber Sardinha, 2019; Fitzsimmons-Doolan, 2014; cf. Biber, 1988), corpus-based discourse analysis research, as well as research in neighboring disciplines, increasingly relies on a similar probabilistic, algorithmic technique called topic modeling (Brookes & McEnery, 2019; DiMaggio, Nag & Blei, 2013; Jaworska & Nanda, 2016; Murakami et al., 2017; Törnberg & Törnberg, 2016a, 2016b). Based on a downsampled research corpus (1,118,454 tokens from 1,257 articles) resulting from analyses of two comprehensive, specialized research (11,656,247 tokens from 16,148 articles) and reference (22,493,804 tokens from 37,227 articles) corpora comprising articles from select Serbian newspapers and magazines published between 2003 and 2008, this study uses WordSmith Tools (Scott, 2014) and SPSS (IBM, 2012), on the one hand, and MALLET (McCullum, 2002), on the other, to compare the relative effectiveness of exploratory factor analysis and topic modeling for identifying large-scale patterns in discourse.

As is well known, both approaches are data-driven and more or less automated and therefore reasonably objective even though they involve subjective decisions at certain points in the analytical process; both offer data amenable to a variety of follow-up synchronic and diachronic analyses, as well as visualization; and both provide an insight into the heteroglossia (multivoicedness) of texts. Exploratory factor analysis, however, is shown to be a more reliable, if also more challenging, methodological solution for large-scale lexical analyses of discursive profiles of unstructured corpora targeting specific research questions (cf. Murakami et al., 2017). Preliminary findings include the following:

- topic modeling (at least using MALLET) is more automated and less cumbersome to run
- although topic modeling is non-deterministic and therefore not entirely replicable, topic models can still be relatively stable and useful
- topic modeling produces results similar and complementary to results produced by exploratory factor analysis, and
- topic modeling can account for a portion of variation unaccounted for by factor analysis, but
- exploratory factor analysis offers a more straightforward, reliable, and replicable path to latent variables (i.e. ‘d’ or thematic discourses), and
- exploratory factor analysis provides a somewhat higher degree of coherence and interpretability because it is based on an analysis of covariation between the collocates of specific core concepts/items (e.g. language) rather than all lexical items in a corpus, so
- topic modeling is probably best used for exploration and validation purposes.

It should also be noted that, although it does not appear to be necessary, lemmatization (and/or stemming) can be useful when dealing with data in highly inflectional languages. Similarly, a three-step approach whereby corpus linguistics and critical discourse analysis are combined to identify (1) “d” (thematic) and (2) “D” (ideological) discourses (cf. Gee, 2010), and (3) language ideologies, as outlined and exemplified in Ajšić (2015, 2016, 2021), is lent further credence. It is finally suggested that, while topic modeling probably should not be used as a standalone method, doubts as to its utility in discourse studies (Brookes & McEnery 2019) may be at least somewhat unwarranted and premature.

## *References*

Ajšić, A. (2015). *Language ideologies, public discourses, and ethnonationalism in the Balkans: A corpus-based study* (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff, AZ. (UMI No. 3705442)

Ajšić, A. (2016). English, “polyglot” politicians and polyglot businessmen: Language ideologies in contemporary Bosnian press. In L. Buckingham (Ed.), *The status of English in Bosnia and Herzegovina* (pp. 159-202). Bristol, Buffalo: Multilingual Matters.

Ajšić, A. (2021). Capturing Herder: A three-step approach to identification of language ideologies using corpus linguistics and critical discourse analysis. *Corpora, 16*(1).

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in UK press. *Discourse & Society, 19*(3), 273-306.

Berber Sardinha, T. (2019). Using multidimensional analysis to detect representations of national identity. In T. Berber Sardinha & M. Veriano Pinto (Eds.), *Multidimensional analysis: Research methods and current issues* (pp. 231-258). London, New York: Bloomsbury Academic.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge: Cambridge University Press.

Brookes, G., & McEnery, A. M. (2019). The utility of topic modelling for discourse studies: a critical evaluation. *Discourse Studies, 1*(21), 3-21.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics 41*, 570-606.

Fitzsimmons-Doolan, S. (2014). Using lexical variables to identify language ideologies in a policy corpus. *Corpora, 9*(1), 57-82.

Gee, P. J. (2010). *An Introduction to discourse analysis: Theory and method.* London: Routledge.

IBM. (2012). Statistical package for the social sciences 21.0 [Computer software]. Retrieved from https://www.ibm.com/ae-en/analytics/spss-statistics-software.

Jaworska, S., & Nanda, A. (2016). Doing well by talking good: A topic modelling-assisted discourse study of corporate social responsibility. *Applied Linguistics, 39*(3), 373-399.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit [Computer software]. Retrieved from http://mallet.cs.umass.edu.

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). What is this corpus about? Using topic modelling to explore a specialised corpus. *Corpora, 12*(2), 243-277.

Partington, A. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) [Special Issue]. *Corpora, 5*(2), 83-108.

Scott, M. (2014). WordSmith tools 6.0 [Computer software]. Stroud: Lexical Analysis Software. Retrieved from https://www.lexically.net/wordsmith/ downloads/.

Subtirelu, N. C. (2013). “English… it’s part of our blood”: Ideologies of language and nation in United States Congressional discourse. *Journal of Sociolinguistics, 17*(1), 37-65.

Subtirelu, N. C. (2015). “She does have an accent but…”: Race and language ideology in students’ evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society, 44*(1), 35-62.

Törnberg, A., & Törnberg, P. (2016a). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse and Society, 27*(4), 401-422.

Törnberg, A., & Törnberg, P. (2016b). Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Discourse, Context and Media, 13*, 132-142.

Vessey, R. (2017). Corpus approaches to language ideology. *Applied Linguistics, 38*(3), 277-296.

**Biri, Ylva**

***University of Helsinki***

**ylva.biri@helsinki.fi**

Type of Contribution: Full Paper

## User positioning online

Process types on Twitter, Tumblr and Reddit

*Abstract*

This study analyzes interest-based online communities in terms of the positions that users attribute to themselves (“self-positioning”). Interest-based communities on social media are established for discussing shared interests: users participate in these communities to connect with each other and to engage in information sharing (Ray et al., 2014). According to positioning theory (Davies & Harré, 1990; Harré & Van Langenhove, 1999), interaction gives rise to situational temporal roles that are adopted and assigned by the interactants. This study describes positions through the kinds of actions a user describes themselves as doing: this kind of implicit self-presentation through a recurring type of action has been found to be more common than explicit self-labels (Bolander & Locher, 2015, studying Facebook).

Here, my primary research question is (1) what positions do users on online communities assign to themselves. As different communities are assumed to have different values and platform settings, the study also asks (2) how do the positions differ between communities. In this study I will focus on the micro-blogs Twitter and Tumblr and the discussion forum Reddit. For a comparison of the influence of different interests, my case study analyzes communities on the topics: *climate change*, *recycling* and *decluttering*. While interest-based communities are sometimes thought of as representing a homogenous type of computer-mediated communication, the positions identified in the data indicates variation that depends on the interest topic and platform conventions.

To study self-positioning expressed through actions, I extracted *I*+(AUXILIARY-VERB)+LEXICAL-VERB as well as *I*+BE+ADJECTIVE clusters from a 3.0-million-word specialized corpus. Collected in November-December 2018-2019 and 2019-2020, the study corpus consists of English-language posts from Twitter, Tumblr and Reddit on the selected interest-topics. On Twitter and Tumblr, the data was collected based on topic-indexing hashtag (#climatechange, #recycling, #decluttering), and on Reddit based on sub-forum (r/climatechange, r/recycling, r/declutter, “r/” being a prefix denoting *sub-reddits*). I analyze the positions and clusters in terms of relative frequency, but also in terms of keyness (log-ratio (Hardie, 2014)) to identify the statistically most salient verbs, which reflect the topic-specific concerns and values of the community. As the register of the study corpus is informal yet quite standard, my reference corpus of 22 million words is composed from subsets of the GloWbE and COCA (spoken, magazine and news) corpora.

The *I*+VERB and *I*+BE+ADJECTIVE clusters are manually categorized by type of action and qualitatively interpreted as positions attributed to community participants. Drawing on the Systemic Functional Linguistics notion of process types (Halliday & Matthiessen, 2013), the main categories identified and analyzed are *material* (doing), *emotive* (feeling) and *verbal-cognitive* (saying, thinking). Self-positioning was found to be least frequent in Twitter-based communities due to a lower frequency of first-person pronouns on Twitter.

Results indicate that differences in the positions have a statistically significant link to the topic as well as the platform. Material processes are especially frequent in decluttering discussions as users narrate their lives through I+VERB constructions. Overall, self-enhancement is a prevalent motive, as qualitative reading of high-keyness verbs shows how material processes are used to position the self in line with the community’s goals through verb clusters such as *I recycle*, *I declutter* and *I donate*. High frequency emotive processes such as *I love* and *I prefer* are typically used for positive stance-taking. Verbal-cognitive processes such as *think*, *know* or *wonder* may serve to epistemically evaluate the discourse. On the other hand, verbs such as *share* or *link* are used in supporting one’s expert position through external evidence or experience. The platform influences positions in terms of both structure and social norms. Verbal-cognitive processes are more frequent on the discussion forum Reddit than on the micro-blogs, likely due to the platform's discussion forum structure and information-seeking nature. Meanwhile, the community's norms play a role as analysis of emotive processes indicates Tumblr's climate change discussion to be the only studied community where the self is positioned as vulnerable though intense negative emotions such as *I am scared*.

The study on positions is limited by its focus on positioning only through verb clusters. However, drawing on sociological and information technological research, the study provides empirical data and working explanations for differences between interest-based online communities. The corpus-based methodology provides a systematic way to explore differences between communities that would not be evident from the obvious topic and platform variables.

## *References*

Bolander, B., & Locher, M. A. (2015). "Peter is a dumb nut": Status updates and reactions to them as 'acts of positioning' in facebook. *Pragmatics*, *25*(1), 99-122.

Davies, B., & Harré, R. (1990). Positioning: The Discursive Production of Selves. *Journal for the Theory of Social Behaviour*, *20*(1), 43-63.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2013). *Halliday's introduction to functional grammar* (4th ed). New York: Routledge:

Hardie, A. (2014). Log Ratio - an informal introduction. ESRC Centre for Corpus Approaches to Social Science (CASS), Blog. 28 April 2014. Retrieved from http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

Harré, R., & Van Langenhove, L. (1999). *Positioning theory: Moral contexts of intentional action*. Malden, MA: Blackwell.

Ray, S., Kim, S. S., & Morris, J. G. (2014). The central role of engagement in online communities. *Information Systems Research*, *25*(3), 528-546.

**Bohmann, Axel**

***University of Freiburg***

**axel.bohmann@anglistik.uni-freiburg.de**

**Müller, Julia**

***University of Freiburg***

**julia.mueller@anglistik.uni-freiburg.de**

**Honkanen, Mirka**

***University of Freiburg***

**mirka.honkanen@anglistik.uni-freiburg.de**

**Neuhausen, Miriam**

***University of Freiburg***

**miriam.neuhausen@anglistik.uni-freiburg.de**

Type of Contribution: Full Paper

## A large-scale diachronic analysis of the English passive alternation

*Abstract*

Alternation between BE- and GET-passives is a well-documented case of linguistic variation in English. It is generally acknowledged that the GET-passive has been increasing in frequency over the past century (Mair, 2006), a development attributed to the combined forces of grammaticalization, colloquialization, and prescriptivist opposition to the BE-passive (Schwarz, 2018), as well as wider cultural change (Anderwald, 2018). The GET-passive is commonly considered to carry a narrower set of connotations than the BE-passive, including adversativity (see Coto-Villalibre, 2015 for a critical

discussion), subject responsibility (e.g. Collins, 1996), and informality (Biber et al., 1999).

Despite the wealth of literature on the GET-passive, however, the variable remains conceptually “fuzzy” (Collins, 1996, p. 44). Rather than presenting a clear-cut case of a single syntactic alternation, a number of constructional contrasts are involved. Among these are the correspondence between active- and passive-voice constructions (e.g. Thompson et al., 2018), the wider set of non-canonical passivization strategies (Alexiadou & Schäfer, 2013), developments in other GET-constructions (Hundt, 2001; Fleisher, 2006), as well as the co-selection preferences individual verbs show for either BE or GET (Rühlemann, 2007). Defining a clear variable context for a statistical analysis of GET- vs. BE-passives is therefore a tricky enterprise that has often required manual inspection of each token in the data.

As a consequence, current insights into the passive alternation are either based on hand-coded analysis of relatively small sets of data or general distributional patterns found in large corpora (e.g. Hundt, 2001; Xiao et al., 2006). Neither of these methods allows for inferential modeling of the large set of hypothesized constraints, let alone their interaction and development over time.

In the present study, we propose an approach that utilizes the quantitative-statistical affordances of the current corpus-linguistic landscape to a fuller extent. Specifically, we trace all instances of variation between GET- and BE-passives across the Corpus of Historical American English (COHA, Davies, 2012), yielding 1,097,898 tokens. The automated extraction procedure necessitates a relaxing of the requirement for strict referential equivalence between variants. However, we address this issue by operationalizing the “centrality” (Quirk et al., 1985) of GET-passives as part of the effect structure of our model. Specifically, for each passivized verb lemma, we include a measure of how often it is pre-modified by *very* and how often it occurs as a complement to copular verbs such as *seem*, *remain*, etc. (Wasow, 1977). Higher scores along both of these measures indicate participles that typically receive an adjectival reading and consequently less central passive constructions.

In order to model the proposed constraints on the GET-passive and their development over time, we further include the following predictors, ordered here by the constraints they relate to:

- Proportional change: we use the year a passive instance is sampled from as a continuous predictor of GET- vs. BE-choice.
- Adversativity: we extract both a verb-level sentiment score (positive, negative, neutral) based on the NRC sentiment lexicon (Mohammad & Turney, 2013) and a sentence level sentiment score (on a five-point scale from clearly negative to clearly positive) based on a combination of two context-sensitive sentiment scoring algorithms (Rinker, 2017; Hutto & Gilbert, 2014).
- Subject responsibility: we manually code the approx. 28,000 most common subjects (96.5% of the data) in terms of animacy on a four-point scale from individual human/animal to collective to metonymically human (body parts) to inanimate.
- Informality: we use both the genre category (fiction, non-fiction books, magazine, newspaper) and a text-level quantitative F-measure (Heylighen & Dewaele, 2002) to operationalize the stylistic context in which a passive token occurs.
- Change in productivity: we create a distributional semantic model of the verb participles in our dataset, based on the corpus in its entirety and using the word2vec algorithm (Mikolov et al., 2013) as implemented in the gensim Python library (Řehůřek & Sojka, 2010) to trace the increasing schematization of the participle slot on the GET passive (cf. the methodology in Perek, 2018). Participles are clustered automatically into interpretable semantic groups by means of Gaussian mixture modelling over the word vector space.

- In addition, we include a number of grammatical predictors, namely: grammatical number of subject, tense, polarity, adverbial pre-modification of the participle, and presence of an agent prepositional phrase headed by.

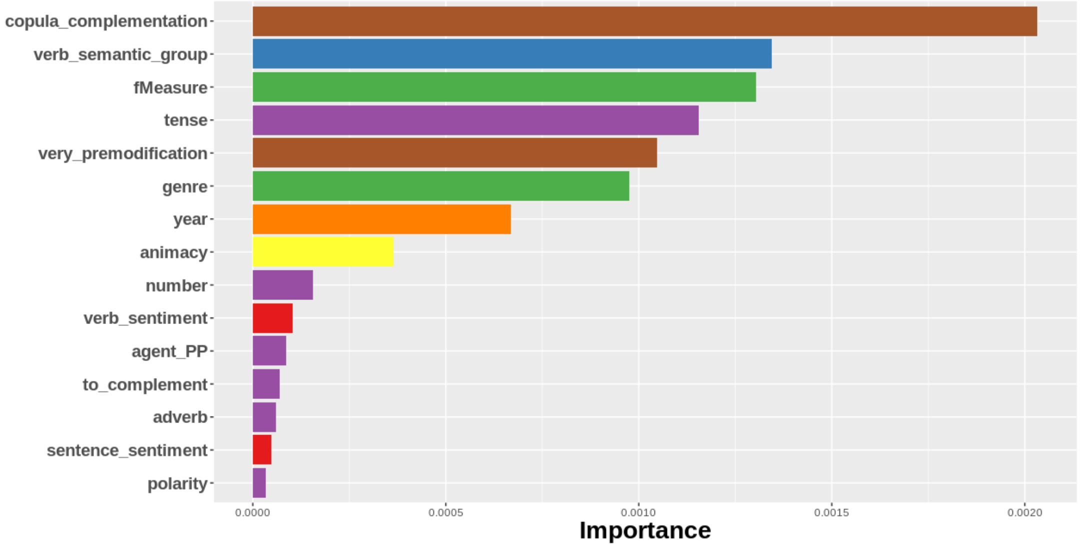


Figure 1: Permutation-based variable importances from a random forest model, weighted to address class imbalance.

We fit two kinds of model to our data: (a) weighted random forests (Breiman, 2001) to account for the skewed ratio between GET and BE in our data; and (b) mixed-effects logistic regressions with random intercepts for individual verb participles and interaction terms between year and all other predictors in order to trace change in constraints over time. Figure 1 shows the variable importance derived from a random forest analysis.

The different predictor variables are color-coded into distinct groups. The first of these, in brown, represents the centrality of a passive token as operationalized through the participle’s propensity to complement a copular verb. Including this variable, along with *very*-premodification in fifth place, therefore helps separate the behavior of central from non-central passives. Next in importance are the different semantic groups in the participle slot, indicated by the blue bar. The importance of this predictor gives post-hoc validation to our choice of distributional semantic modelling. Measures of formality can be found in green, ranked third (f-measure) and sixth (genre) in terms of their importance. Of further note are the influence of tense/aspect (purple bar in fourth place), of year (orange, seventh), and subject animacy (yellow, eighth). Sentiment and further syntactic information, such as the presence or absence of an agent-PP with by, are comparatively less relevant. In sum, Figure 1 makes a compelling argument that aspects of the verb participle - its semantic group membership as well as passive centrality measures - are by far more important than syntactic properties in predicting the choice of the GET-passive. This is all the more significant since in the extant literature it is often the latter that are emphasized over the former (although see Rühlemann, 2007).

Naturally, the ordering of variable importance does not give any information about linear relationships between predictors and the outcome variable. The process of fitting and comparing mixed-effects logistic regression models on a data set of this size is computationally expensive and can take several days per model. For this reason, our search for an optimal linear model is still ongoing. Tentative results indicate that the maximal model fit thus far prioritizes grammatical and stylistic predictors. Finding a unified account for results from both random forest and linear modeling will be an important task in the future.

Our analysis at this point indicates a need to reconsider some assumptions about the mechanisms involved in the passive alternation, namely: the prioritization of grammatical constraints and the comparative neglect of verb participle semantics. Beyond the relevance of the concrete empirical results, this study demonstrates the utility of principled computational methods for modelling semantic influences on syntactic variation.

## *References*

Bohmann, A. (2019). *Variation in English worldwide: Registers and global varieties*. Cambridge: Cambridge University Press.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide, 36*(1), 1-28.

Gonçalves, B., Loureiro-Porto, L., Ramasco, J. J., & Sánchez, D. (2018). Mapping the Americanization of English in space and time. *PloS ONE, 13*(5): e0197741.

Kachru, B. (1985). Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11-30). Cambridge: Cambridge University Press.

Mikolov T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*: 1301.3781.

Mukherjee, J., & Bernaisch, T. (2015). Cultural keywords in context: A pilot study of linguistic acculturation in South Asian Englishes. In P. Collins (Ed.), *Grammatical change in English world-wide* (pp. 411-436). Amsterdam, Philadelphia: John Benjamins.

Rehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceeding of the LREC Workshop on New Challenges for NLP Frameworks*.

Schneider, E. W. (2007). *Postcolonial English: Varieties around the world.* Cambridge, New York: Cambridge University Press.

Seoane E., & Suárez-Gómez, C. (Eds.). 2016. *World Englishes: New theoretical and methodological considerations*. Amsterdam; Philadelphia: John Benjamins.

**Bohmann, Axel**

***University of Freiburg***

**axel.bohmann@anglistik.uni-freiburg.de**

Type of Contribution: Full Paper

## Varieties of English worldwide

A lexical-semantic perspective

*Abstract*

Corpus research on relationships between varieties of English to date focuses predominantly on morpho-syntactic variation, either through comparative single-feature studies (e.g. contributions in Seoane & Suárez-Gómez, 2016) or through an aggregate approach (Bohmann, 2019). A lexical perspective has so far only been applied in a few studies, focusing either on cultural keywords (Mukherjee & Bernaisch, 2015) or on a small set of words with clear varietal distribution between British and American English such as *aubergine* and *eggplant* (Gonçalves et al., 2018).

The present study addresses lexical differences in World Englishes at both a finer level of detail and with a greater breadth of coverage of individual words. To this purpose, a distributional analysis is performed for the 1000 most frequent word lemmas in the Corpus of Global Web-based English (GloWbE, Davies & Fuchs, 2015). In a distributional model, a word's meaning is expressed in terms of its distance from other words in a high-dimensional vector space whose dimensions represent the collocation behavior of each word. Drawing on the word2vec algorithm (Mikolov et al., 2013) as implemented in Python's gensim library (Rehůřek & Sojka, 2010), such a vector-semantic model is constructed for each national sub-corpus of GloWbE.

Each word in each variety is then represented as a vector of distances to all other words that occur at least 1,000 times in GloWbE and at least once in each variety. This is done to ensure comparability across varieties. For any given word, the distance between two varieties is calculated as the cosine

distance of the word’s vector representations in the two varieties. With this method, word-level distance matrices can be calculated for the 20 varieties in GloWbE, which in turn can be represented as phylogenetic trees to show inter-varietal difference/similarity in respect to this particular word. Figure 1 depicts such visualizations for the four most common words in the corpus.

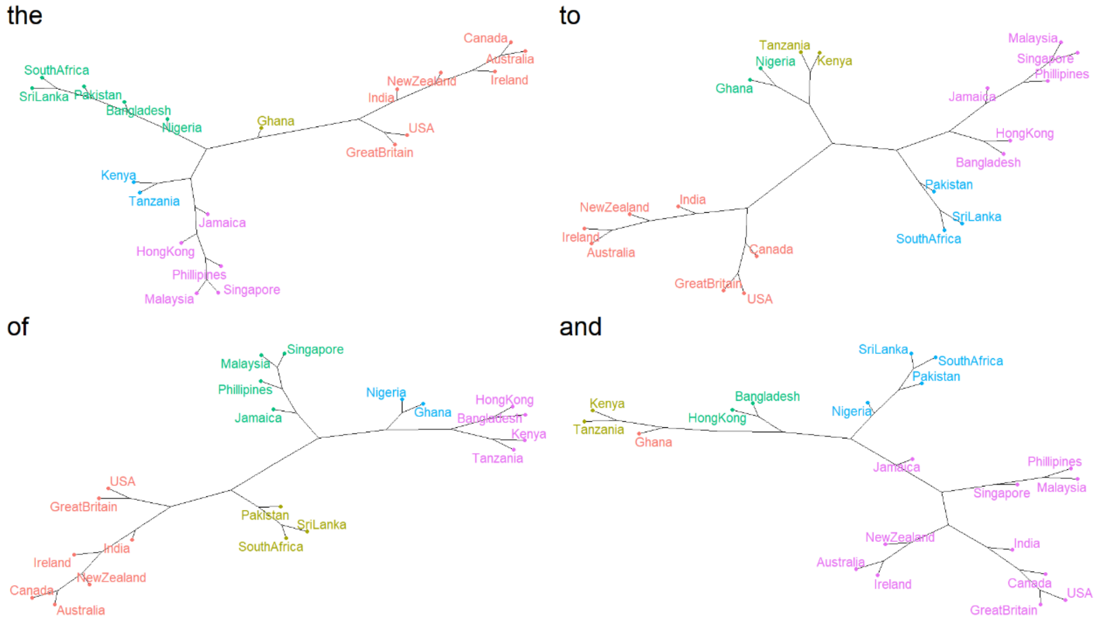


Figure 2: Relations among varieties in GloWbE based on the four most frequent words.

All four tree diagrams consistently cluster Inner Circle (Kachru, 1985) / Phase 5 (Scheider, 2007) varieties together. Beyond this, regional patterns can be identified in all four graphs; however, there is considerable word-level variation as to the precise shape of regional differentiation.

In order to construct a more robust picture of general inter-varietal relationships in word usage, the individual distance matrices for words are weighted to represent their token frequency in the corpus on the whole and added together. After aggregating distances for the most frequent 1,000 words, the tree diagrams reach a point of stability. This can be interpreted as evidence for the empirical robustness of the distributional model. The corresponding aggregate tree diagram is shown in Figure 2.

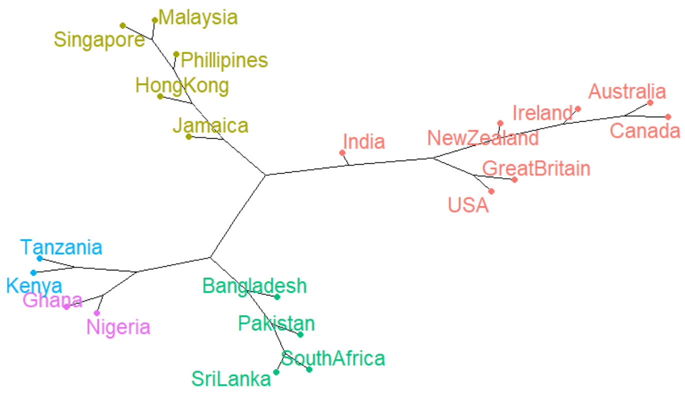


Figure 3: Relations among varieties in GloWbE based on the most frequent 1,000 words.

Figure 2 confirms the impression of an important split between Inner Circle / Phase 5 and all other varieties. The latter further show a very clear pattern of regional differentiation into a Southeast Asian, an Asian, and an African cluster. Three varieties are positioned in ways that require further explanation: (a) South Africa tends to cluster with the South Asian varieties, a fact that may in part be explained in terms of migration history from India to South Africa; (b) Jamaica is closest to the Southeast Asian varieties. Being the only variety in a (post-)creole continuum situation represented in GloWbE is likely to give Jamaican English a special status; (c) finally, India is found among the Inner Circle countries, perhaps because of the close adherence of Indian English to British norms. It has to be stressed that in all three cases, these explanations are tentative at present and require further corroboration.

Methodologically, it can be stated with confidence that a distributional analysis of World Englishes yields plausible patterns by and large, reflecting both differences predictable from the existing models as well as a strong regional signal. The latter is noteworthy in particular because the model presented here was produced on the basis of the most common shared words and thus does not capitalize on individual, locally specific terms. A basic tension between American and British influence cannot be confirmed as playing a significant role in the data at hand, as the USA and Great Britain are more similar to each other than to any other variety.

At this point, the method is largely exploratory. More work needs to be done to identify what drives the large-scale patterns visible in Figures 1 and 2 above. To that end, future steps will include clustering of the distance matrices for individual words, factor analysis of the dimensions of inter-varietal differentiation, and qualitative analysis of lexical items identified as particularly distinctive from the above two approaches.

### *References*

Bohmann, A. (2019). *Variation in English worldwide: Registers and global varieties*. Cambridge: Cambridge University Press.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide, 36*(1), 1-28.

Gonçalves, B., Loureiro-Porto, L., Ramasco, J. J., & Sánchez, D. (2018). Mapping the Americanization of English in space and time. *PloS ONE, 13*(5): e0197741.

Kachru, B. (1985). Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the World: Teaching and learning the language and literatures*, (pp. 11-30). Cambridge, New York: Cambridge University Press.

Mikolov T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*: 1301.3781.

Mukherjee, J., & Bernaisch T. (2015). Cultural keywords in context: A pilot study of linguistic acculturation in South Asian Englishes. In P. Collins (Ed.) *Grammatical change in English world-wide* (pp. 411-436). Amsterdam, Philadelphia: John Benjamins.

Rehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. Proceedings of LREC Workshop on New Challenges for NLP Frameworks.

Schneider, E. W. (2007). *Postcolonial English: Varieties around the world.* Cambridge, New York: Cambridge University Press.

Seoane E., & Suárez-Gómez, C. (Eds.). (2016). *World Englishes: New theoretical and methodological considerations*. Amsterdam, Philadelphia: John Benjamins.

**Bourgeois, Samuel**

***University of Neuchâtel***

**samuel.bourgeois@unine.ch**

Type of Contribution: Full Paper

## “Creed was, well, Creed.”

## On shifts in the use of *well* and *like* in American journalism from the 1990s onward

*Abstract*

This paper investigates the colloquial and (inter)subjective non-dialogical uses of the discourse markers (DMs) *well*, and *like* in American journalistic texts based on corpus data from the Corpus of Historical American English (COHA) (Davies, 2010) and the News on the Web corpus (NOW) (Davies, 2013). In particular, it demonstrates that such uses are a recent phenomenon, which have risen rapidly in frequency since the 1990s onward. In journalism, these DMs are developing genre specific functions that resemble their textual oral functions of conducting word-searches or performing self-corrections but are, at the same time, adapted to suit the rhetorical needs of writing. Moreover, these DMs are increasingly used to indicate important lexical choices while also expressing their (inter)subjective stance (c.f. Du Bois, 2007) (see examples 1 and 2).

(1) Now, it’s not all bad — her boyfriend loves comic books; the best baes always will — but Sabrina double-lifes in the Church of Night, a coven that worships, **um**, **like**, **well**, the Devil. (NOW: Entertainment Weekly, 2018)

(2) With the character set for his solo cinematic debut in a few years, it’s about time that DC Entertainment’s Aquaman got some recognition for his constant, under the radar (**well**, under the sonar, technically speaking), onscreen appeal over the last few decades. (NOW: Hollywood Reporter, 2014)

This paper combines quantitative and qualitative analysis to investigate these DMs in written journalism. Using the Variability-based Neighbor Clustering (VNC) algorithm (c.f. Gries & Hilpert, 2008; Hilpert, 2013) to highlight the diachronic stages of the usage of these two DMs in writing, this paper demonstrates that they show a remarkable increase in usage in the journalistic sub-corpora of the COHA starting in the 1990s. These notable surges in usage in the 1990s are the result of their continued increased usage in transcribed speech but most notably the sudden appearance of their usage in non-dialogical rhetorical functions. *Well* and *like* continue to be increasingly used in journalistic texts in the 2000s and even into the 2010s data, which comes from NOW corpus. These developments contribute further insights to what we already know about change occurring in journalistic writing (c.f. Westin & Geisler, 2002; Mair, 2006). For example, Leech et al. (2009, p. 239) state that the increasing colloquial style of journalism gives it "a kind of spontaneous directness which (though often contrived) is clearly supposed to inject into journalistic discourse some of the immediacy of oral communication". While these studies indicate an increased usage of quotations and the preference of authors to use more informal constructions or conventions when a more formal option is also available (Mair, 2006, pp. 188-189), this paper shows how these colloquial DMs are adapted into journalistic writing in ways that are unique to writing.

The qualitative look at *well* and *like* investigates the individual ways in which these two DMs are used in non-dialogical contexts. Though these two DMs are used in ways that resemble the author actively thinking about what (s)he wants to say, or correcting oneself, the usage of the DMs in journalistic prose is found to encode the author's (inter)subjective stances toward the lexical items (or upcoming information) being highlighted. Furthermore, the fact that these DMs are increasingly adopted into journalistic prose in functions that encode the author's subjective and intersubjective stance dovetails with observations made by Traugott & Dasher (2002) and Traugott (2010) in terms of the (inter)subjectification of DMs in general where their meanings shift from non/less subjective to subjective and then intersubjective. However, unlike prior observations on the development of these DMs, the increased (inter-) subjectification in journalistic prose comes from the expansion of the textual functions that primarily occur in the sentence medial positions and involve highlighting the rhetorical use of lexical items specifically.

One factor that has undoubtedly contributed to the development of these specifically journalistic functions of *well* and *like* is the salience of their use. Considering the fact that texts are edited before being published, the use of DMs marking word-searching or self-correction in journalistic prose have no tangible reason to be used. The salience of the DMs when used in writing in ways that resemble such uses, however, does explain their usefulness in this medium, especially in terms of making the texts seem more interpersonal and oral-like. Using these two DMs in writing in functions that resemble word-searches and self-corrections allow the author to mark his/her (inter)subjective attitudes in new and innovative ways that are highly salient, colloquial-like as well as specific to the journalistic style of writing.

In addition to these observations on the development of (inter)subjective and genre specific DM functions, this work also highlights the importance of verifying from what type of articles one finds these characteristics of colloquialization. For example, in contrast to Rühlemann & Hilpert's (2017) analysis of *well* in the TIME Corpus (Davies, 2007), this study has taken note of the types of articles in which the non-dialogical uses of these DMs have been used. Furthermore, in the case of the NOW data, it also notes whether the source is from traditional print sources or web-only sources. From their sudden uses in non-dialogical functions from the 1990s to today, the usage of these two DMs is mostly confined to the informal sub-genres of journalism. This is especially the case in entertainment news, sports and op-eds/columns articles.

Finally, the timing of these developments indicates a later wave of change to journalism that has, so far, been under-discussed in colloquialization studies. To my knowledge only Rühlemann & Hilpert (2017)'s study of *well* and Tottie's (2017) investigation of *uh*, *um* and *er* describe similar jumps in the usage of colloquial features in journalistic prose starting in the 1990s. The changes in how DMs are used in journalistic writing are influenced by many different factors. This includes, of course, the changes that have already been observed in terms of colloquialization. This shift in the writing style of journalism is likely further encouraged at the end of the twentieth century thanks to the vernacular characteristics of the online writing style that are increasingly popular and interconnected in certain informal areas of journalism, like for instance, entertainment news (e.g. Barton & Lee, 2013).

### *References*

Aijmer, K. (2013). *Understanding pragmatic markers: A variational pragmatic approach*. Edinburgh: Edinburgh University Press.

Barton, D. & Lee, C. (2013). *Language online: Investigating digital texts and practices*. London: Routledge.

Davies, M. (2010). The corpus of historical American English (COHA): 400 million words, 1810-2009. Retrieved from https://www.english-corpora.org/coha/

Davies, M. (2013). Corpus of news on the web (NOW): 3+ billion words from 20 countries, updated every day. Retrieved from https://www.english-corpora.org/now/

Du Bois, J. W. (2007). The stance triangle. In R. Englebretson (Ed.), *Stancetaking in discourse: Subjectivity, evolution, interaction* (pp. 139-182). Amsterdam: John Benjamins.

Gries, S. Th. & Hilpert, M. (2008). The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora, 3*(1), 59-81.

Hilpert. M. (2013). *Constructional change in English: Developments in allomorphy, word-formation, and syntax*. Cambridge: Cambridge University Press.

Leech, G. & Hundt, M. & Mair, C. & Smith, N. (2009). *Change in contemporary English: A grammatical study.* Cambridge: Cambridge University Press.

Mair, C. (2006). *Twentieth-Century English: History, variation and standardization*. Cambridge: Cambridge University Press.

Rühlemann, C. & Hilpert, M. (2017). Colloquialization in journalistic writing: The case of inserts with a focus on well. *Journal of Historical Pragmatics, 18*(1), 104-135.

Tottie, G. (2017). From pause to word: uh, um and er in written American English. *Language and Linguistics, 23*(1), 105-130.

Traugott, E. C. (2010). (Inter)subjectivity and (inter)subjectification: A reassessment. In K. Davidse, L. Vandelanotte & H. Cuyckens (Eds.), *Subjectification, Intersubjectification and Grammaticalization* (pp. 29-74). Berlin, Boston: De Gruyter Mouton.

Traugott, E. C. & Dasher, R.B. (2002). *Regularity in semantic change*. Cambridge: Cambridge University Press.

Westin, I. & Geisler, C. (2002). A multi-dimensional study of diachronic variation in British newspaper editorials. *ICAME Journal 26*, 133-152.

**Brookes, Gavin**

***Lancaster University***

**g.j.brookes@lancaster.ac.uk**

Type of Contribution: Full Paper

## Gendered discourses of obesity

A corpus-based study of the British Press

*Abstract*

This talk examines how press representations of obesity intersect with discourses around gender in the context of the British press. The World Health Organization (2020) defines *obesity* as ‘abnormal or excessive fat accumulation that may impair health’. *Gendered discourses* are taken to be those practices - including linguistic practices - which establish ‘boundaries of social practice through which appropriate gendered behaviour is regulated’, providing parameters through which people are “represented or expected to behave *in particular gendered ways*” (Sunderland, 2004, p. 21). This talk analyzes how men and women are represented with respect to obesity, linking these representations to wider discourses around gender at the societal level. In this sense, the analysis will explore how obesity interacts both with sex and gender in the press.

The data examined in this talk is a sample of weight loss narratives taken from a large corpus (approx. 44,000 articles, 36 million words) of British national press articles mentioning *obese* or *obesity* between 2008 and 2017. By *weight loss narratives*, I refer to articles published by newspapers, particularly tabloids, that report on the weight loss of an individual, but sometimes couples. They can be about celebrities or ordinary people and they tend to be multimodal, often featuring ‘before-and-after’ images. They can be presented either in third or first-person, but in either case are carefully selected and crafted by the editor for the newspapers’ imagined readership (Bell, 1991). Thus, while they reflect the experiences of individuals, they can’t be treated as subjective disclosures

or accounts of weight loss. To identify weight loss narratives and isolate a sample for analysis, I read through tabloid articles, taken at random, and extracted the first 100 weight loss stories about women and the first 100 stories about men that I encountered, and stored these, in their entirety, as two samples. The search was restricted to just the tabloids as prior analysis has shown these types of articles to be particularly characteristic of the tabloids rather than the broadsheets. The resulting samples represented a cross-section of all but one of the national tabloids (*Morning Star*) and were fairly evenly matched in terms of size, with the men's narratives containing 106,269 words and the women's reaching 100,187. Analysis is driven by a keyword comparison of these two samples of articles against each other, with the resulting keywords indicating aspects of the representation of the men and women in the narratives that were characteristic of each. Keywords are analyzed in context and interpreted in terms of the representations they reflect which are, in turn, interpreted as constituting gendered discourses.

Taken together, the identified gendered discourses provide evidence that the press recycles and reinforces a restricted set of prominent and long-standing ideologies around femininity and masculinity. As well as being reported on much more than men, women are consistently represented in terms of relational aspects of their identity, with particular focus on their roles in, and responsibilities for, caring for others, including their (unborn) children and spouses. The relational identities afforded to the women also manifests in an aesthetic focus on their bodies, weight and weight loss, whereby both their desire to lose weight and their weight loss results are presented in terms of how they appear to others, for instance in wedding dresses or bikinis, prompting evaluations of them in the articles as 'beautiful' or 'stunning'. Representations of women also foreground the emotional aspects of their weight loss and embodied experience, with obesity presented as causing embarrassment and even depression, the process of weight loss characterized in terms of difficulty and desperation, and the results of weight loss leading to increased confidence and other more positive emotions.

Men, by contrast, are portrayed in ways that backgrounded their emotions. Instead, their decisions to lose weight were construed as logical and life-preserving, as motivated by the threat of disease or even death that was posed by their obesity. Although the word *diet* is key in the men's narratives relative to the women's, their weight loss actions are frequently portrayed as *not* being diets, in some cases lexicalized instead as 'experiments'. The diets that men *do* engage in are marked or unusual and frequently preserve the consumption of foodstuffs that have become symbolic of hegemonic masculinity, such as alcohol and red meat.

The discourses identified thus offer a relatively narrow range of gender roles and experiences which likely fail to capture the complexity and variability of contemporary gender identities, including the ways in which these intersect with notions and experiences of health. This trend is interpreted as reflecting the tabloids' attempts to provide 'traditional' views of gender that they perceive to be consonant with their older 'imagined' readerships (Bell, 1991). Given that such traditional views of gender and the stereotypes about men and women that they give rise to have been argued to have detrimental impacts on the health of women and men alike, it is argued that the public would be better served by media coverage that challenges such norms or which at least incorporates a wider range of gendered subject positions that better reflect the range of gendered identities within contemporary society.

### *References*

Bell, A. (1991). *Language of news media*. New York: John Wiley and Sons.

Sunderland, J. (2004). *Gendered discourses*. Basingstoke: Palgrave Macmillan.

World Health Organization. (2020). Obesity and overweight. Online. Retrieved from: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

**Brůhová, Gabriela**

***Charles University***

**gabriela.bruhova@ff.cuni.cz**

**Vašků, Kateřina**

***Charles University***

**katerina.vasku@ff.cuni.cz**

Type of Contribution: Work-in-Progress Report

# *Of*-structures in L2 academic English

## Lexico-semantic features of syntactic complexity

*Abstract*

The study is a part of an ongoing project focusing on phraseology of Czech users of advanced academic English. The present paper explores the use of sequences containing the preposition *of* by Czech advanced learners of academic English. The analysis is based on two corpora: a corpus of L2 novice academic writing VESPA-CZ and an L1 corpus of thematically congruent papers (namely papers on English literature) published in academic journals. Since the preposition *of* is significantly less frequent in L2 novice academic texts written by Czech speakers in comparison with L1 professional writers, the aim of this paper is to investigate to what extent and in what ways the use of sequences containing *of* differs in the two corpora, focusing on their structural and lexico-semantic features. As pointed out by Groom (2010, p. 63), "*of* constitutes an excellent test-bed for the claim that closed-class keywords are tractable to qualitative semantic analysis".

Previous research has suggested that language produced by advanced L2 speakers can be influenced by their limited lexical and phraseological choices (Granger, 2017). Hence, the present study intends to contribute towards developing a phraseology-informed approach to language instruction. Complexity, as one of the three components of L2 proficiency, is usually

described as “the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2” (Bulté & Housen, 2012, p. 2). Although the majority of research studies on syntactic complexity have discussed the formal properties of language, Ryshina-Pankova argues that “the focus on meaning is inseparable from the functions complexity serves in various contexts of human verbal interaction” (2015, p. 60).

Our study is carried out on the basis of 600 semantic sequences (cf. Hunston, 2008) containing *of* (300 instances from each corpus). Following (and elaborating upon) Groom’s (2010) classification, we have analyzed the sequences both structurally and semantically. The structural analysis revealed that there are not significant differences in the distribution of the main formal categories in the L1 and L2 corpora. The most frequent category in both corpora is **n *of* n**, the remaining categories, **prep n**, **adj *of* n**, and **v *of* n** are represented only marginally.

| **structural type** | **L2 corpus** | | **L1 corpus** | | |
|---|---|---|---|---|---|
| **n *of* n** | 264 | 88% | 273 | 91% | *similarities of structure; the problems of identity* |
| **prep n** | 14 | 4.7% | 15 | 5% | *because of outer pressure; in the course of the play* |
| **adj *of* n** | 13 | 4.3% | 4 | 1.3% | *full of racial feeling; reminiscent of a constant dance* |
| **v *of* n** | 9 | 3% | 8 | 2.7% | *hears of his son’s death; consists of impersonation* |
| **total** | 300 | 100% | 300 | 100% | |

Table 1: Structural analysis.

Next, the most prominent structural type, namely **n *of* n**, was subjected to semantic sequence analysis. Altogether, eleven semantic sequences (i.e., “repeated sequences of underlying meanings”, cf. Groom, 2010, p. 65) were identified, which demonstrates “enormous combinatory potential of *of* in English” (ibid.). The six most frequent types are listed below:

(1) PROPERTY + *of* + PHENOMENON (81 instances in the L2 corpus, 85 in the L1 corpus):
*a unity of his argument* (L2)*, morality of allegorical poetry* (L1)

(2) PROCESS + *of* + OBJECT (61 instances in L2, 77 in L1):
*his completion of his plan* (L2), *these simple translations of rhetorical principle*s (L1)

(3) SIGNALLING NOUN + *of* + SPECIFICATION (51 instances in L2, 42 in L1):
*the character of Margery* (L2), *the idea of carving in living flesh* (L1)

(4) QUANTITY + *of* + PHENOMENON (26 instances in L2, 18 in L1):
*amount of information* (L2), *the majority of narrative roles* (L1)

(5) PROCESS + *of* + ACTOR (20 instances in L2, 10 in L1):
*the passing of time* (L2), *conflation of human and artificial traits* (L1)

(6) TEXT + *of* + CONTENT (3 instances in L2, 13 in L1):
*the list of crimes* (L2), *lines of criticism* (L1)

The minor types of semantic sequences, which were represented by less than 10 instances in both corpora, include: ACTOR + *of* + OBJECT (*a reader of poetry*, L2), AUTHORITY + *of* + DOMAIN (*sovereign of Egypt;* L2), TIME + *of* + PHENOMENON (*the first day of July;* L1), PHENOMENON + *of* + PROPERTY (*the points of difference;* L1), TEXT + *of* + ACTOR (*works of Early Modern poets*, L2).

The overall results suggest that, apart from the most common semantic sequence PROPERTY + *of* + PHENOMENON, a large number of instances (27% in L2 and 29% in L1) participate in semantic sequences expressing nominalized processes, which is a typical feature of academic texts. However, there is a difference between the two corpora in the fact that the preference for PROCESS + *of* + OBJECT over PROCESS + *of* + ACTOR in L1 is more significant than in L2 where the latter sequence is overrepresented. The study has also revealed a new type of semantic sequence, SIGNALLING NOUN + *of* + SPECIFICATION, which seems to be characteristic of academic discourse (cf. Flowerdew, 2003).

The semantic analysis has shown that L2 novice academic writers are able to use *of* to form structures which convey a wide range of meanings, similarly to L1 users. However, it appears that the minor types of semantic sequences are used less frequently by L2 writers, which suggests that they are less familiar with them. Suggestions for further research of lexico-semantic features of syntactic complexity include the focus on sophistication of the noun on the left of the preposition *of*, or the study of other, more semantically specific, closed-class keywords.

### *References*

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. Amsterdam, Philadelphia: John Benjamins.

Flowerdew, J. (2003). Signaling nouns in discourse. *English for Specific Purposes, 22*(4), 329-346.

Granger, S. (2017). Academic phraseology: A key ingredient in successful L2 academic literacy. *Oslo Studies in Language, 9*(3), 9-27.

Groom, N. (2010). Closed-class keywords and corpus-driven discourse analysis. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 59-78). Amsterdam: John Benjamins.

Hunston, S. (2008). Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics, 13*(3), 271-295.

Ryshina-Pankova, M. (2015). A meaning-based approach to the study of complexity in L2 writing: The case of grammatical metaphor. *Journal of Second Language Writing, 29*, 51-63.

**Büker, Jonathan**

***TU Dortmund***

**jonathan.bueker@tu-dortmund.de**

**Dolberg, Florian**

***TU Dortmund***

**florian.dolberg@udo.edu**

Type of Contribution: Full Paper

# CLaSS

## Corpus Linguistics and School Settings

*Abstract*

The language classroom is the prime locus for the Next Generation to early and meaningfully experience Interdisciplinarity of and Transfer between corpus linguistics (CL) and language learning. While decades of research in CL for language teaching (cf. e.g. Ghadessy et al., 2001; O'Keeffe et al., 2007; Timmis, 2016) yielded numerous positive impulses, actual application of CL-methodology in school is still rare (cf. e.g. Reppen, 2010; Zareva, 2017; Vincent & Nesi, 2018).

We argue that CL is a powerful tool for classroom work, because it is a versatile, computer-based, and scientific approach to language. Students themselves applying scientific methods as a vehicle to learn about language can foment their scientific, media, and IT-competence, alongside gaining language awareness, proficiency, and communicative competence. Simple methodologies, freeware, and objects of investigation that form a meaningful part of students' every-day life are crucial to achieve this.

A significant part of both CL and language learning/communicative competence pertains to register/genre/text-type differences (cf. e.g. Henry & Roseberry, 2001 or Lee, 2001 for definitions). Traditionally, these differences are taught deductively in top-down fashion through materials which often fail to

spark 10th-to-12th-graders' interest. Instead, we propose an inductive, hands-on approach: students build their own corpora from material of their own choosing, then analyze these by means of simple frequency- and/or keyword-lists.

To test this idea, we compiled three corpora of music lyrics, representing Pop, HipHop and Metal. During our 'laboratory test', we encountered several issues which are important to keep in mind when working with high-school students:

We decided to use the 25 most popular songs from each of said three music styles as a placeholder and benchmark. To this end, we took the top 25 HipHop and Pop songs from the respective Billboard charts. However, Billboard does not maintain a chart list for Metal. To keep the sources comparable (an important point to drive home when doing science with beginners), we supplemented the Billboard song lists with the 25 top entries in the iTunes Metal charts. The lyrics themselves we copy-pasted from websites such as https://www.lyrics.com/, https://www.azlyrics.com/, http://www.songlyrics.com/ or https://genius.com/. Another important issue is to filter out cover versions if these cross genre boundaries: for example, the original lyrics used in a Metal band's cover version of a Pop song still represent and instantiate Pop song lyrics and would hence contaminate a corpus of Metal lyrics. These and some other issues are easily solved by simply eyeballing the data. Eyeballing is also an important heuristic to be taught to beginners and, for small corpora such as in this case, a good alternative to using a scraper (cf. Brett & Pinna, 2019, p. 313). The final issue to report here concerns which software to use for working with the completed corpora. We selected AntConc (Anthony, 2019), because it is free, lightweight, and is able to run on all major operating systems. Moreover, it is relatively easy to use while offering quite a sophisticated feature set.

Word lists generated from our three small corpora yielded several interesting results: As expected, HipHop displays the highest token count per song, meaning HipHop songs tend to have longer lyrics than Pop or Metal songs. Somewhat unexpectedly, Metal displays the highest average type/token ratio, i.e. Metal songs tend to exhibit richer, more varied vocabulary than Pop or HipHop songs. The frequency of certain tokens also provide useful insight: *I* is by a comfortable margin the most frequent word in HipHop and Pop, whereas in Metal it is *the*. Taking this frequency rating of function words as a simple heuristic to indicate the predominant function of language (following Jakobson's (1960) classic model), this implies that HipHop and Pop lyrics are mostly expressive, while Metal lyrics are chiefly referential.

The ranking of content words reveals some information about vocabulary typical for the respective genre, and hence macro-topics that are prevalent in the associated (sub-)culture: In the HipHop corpus, expletives like *shit*, *nigga*, and *bitch* rank very high compared to the other corpora (if they contain profanities at all). Words like *money* and *need* also rank high in HipHop only. On this basis, central topics in HipHop culture are the breaking of (verbal) taboos and concern with material possessions. In the Pop corpus, words of affection (e.g. *love*, *baby*) rank high in the frequency list - unsurprising given that matters of the heart are central to Pop's raison d'être. The Metal corpus is the only one in which adverbs of time (*again, never*, *now*) rank relatively high. This finding is less straightforward to interpret: it might reflect Metal lyrics being chiefly referential (s.a.), and/or it might indicate that the temporal (and hence finite) nature of the human condition is a central topic in Metal culture, or something else entirely. At the very least, it is well-suited for practicing hypothesis construction - another central aspect of science and science education.

These basic findings already clearly demonstrate: Since word choice and frequency are key differentiators of register/genre/text-type differences, corpora of music lyrics are useful for teaching awareness thereof. Our results moreover illustrate that simple CL in the language classroom can well be a viable means for language teaching and learning, complementing more traditional methods. The methodology thus established serves as a springboard for further CL-driven exploration of register/genre/text-type, but also for CL-based approaches to e.g. phraseology, vocabulary, or language variation and varieties, with a view to anchor (corpus) linguistics in school curricula.

## *References*

Anthony, L. (2019). *AntConc.* Retrieved from: https://www.laurenceanthony.net/software/ antconc/.

Apple Inc. (n.d.) *iTunes*. https://www.apple.com/itunes/.

Billboard. (n.d.) *Billboard Charts*. https://www.billboard.com/charts/.

Brett, D., & Pinna, A. (2018). Words (don’t come easy): The automatic retrieval and analysis of popular song lyrics. In C. Suhr, T. Nevalainen & I. Taavitsainen (Eds.), *From data to evidence in English language research* (pp. 307-325). Leiden: Brill.

Ghadessy, M., Henry, A., & Roseberry, R.L. (Eds.). 2001. *Small corpus studies and ELT: Theory and practice*. Amsterdam: John Benjamins.

Henry, A., & Roseberry, R. L. (2001). Using a small corpus to obtain data for teaching a genre. In M. Ghadessy, A. Henry & R. L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 93-134). Amsterdam: John Benjamins.

Jakobson, R. (1960). Closing statement: Linguistics and poetics. In T. A. Sebeok (Ed.), *Style in language* (pp. 339-377). The Technology Press of Massachusetts Institute of Technology/John Wiley & Sons.

Lee, D. Y. W. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology, 5*(3), 37-72.

O’Keeffe, A., McCarthy, M., & Carter, R. (2007*). From corpus to classroom*. Cambridge: Cambridge University Press.

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

Timmis, I. (2016). *Corpus linguistics for ELT: Research and practice*. New York: Routledge.

Vincent, B., & Nesi, H. (2018). The BAWE Quicklinks Project: A New DDL Resource for University Students. *Lidil,* 58.

Zareva, A. (2017). Incorporating corpus literacy skills into TESOL teacher training. *ELT Journal, 71*(1), 69-79.

**Candarli, Duygu**

***University of Dundee***

**dcandarli001@dundee.ac.uk**

**Dang, Yen**

***University of Leeds***

**T.N.Y.Dang@leeds.ac.uk**

Type of Contribution: Full Paper

# The functional and lexical profiles of online academic forum posts

Combining corpus methods with qualitative discourse analysis

*Abstract*

This study combines quantitative corpus methods with qualitative discourse analysis (see Egbert & Baker, 2019) to investigate online academic forum posts written by first language (L1) and second language (L2) writers at a university in the UK. Academic forum posts refer to students' posts written for assessment purposes in a virtual learning environment. The audience of these forum posts were primarily their peers and secondarily lecturers/graders. We examined a corpus of online academic forum posts, which consisted of 240 successful texts that received high grades, written by L1 English and L1 Chinese postgraduate students studying for a master's degree in education. We addressed the following research questions:

(1) What are the communicative functions of online academic forum posts and how do they vary across L1 backgrounds and genres?

(2) How many words are needed to reach 95% and 98% coverage of the most frequent communicative functions of these posts?

(3) What are the typical multi-word sequences of the most frequent communicative functions of these posts?

In order to answer the first research question, we used Swales' genre analysis framework (1990, 2004) and qualitatively coded all the forum posts according to their communicative functions, using NVivo 12 (QSR International, 2018). Based on our qualitative analysis, we propose a new functional taxonomy of online academic forum posts that are written for assessment purposes and extend the functional taxonomy that was used in previous studies (e.g. Ädel, 2011; Li & Kim, 2016). To answer the second research question, we divided the corpus of online academic forum posts into sub-corpora according to their communicative functions. Then, we used RANGE (Heatley, Nation, & Coxhead, 2002) and AntConc (Anthony, 2019) to identify the single words and multi-word sequences that frequently occur in each sub-corpus. To answer the third research question, we compared the items identified in the above step. The results reveal great variations in terms of the lexical coverage and multi-word sequences across the communicative functions of the forum posts. This suggests the importance of qualitative functional analysis combined with quantitative corpus methods. Practical implications for teaching academic writing and online academic literacies are also discussed.

*References*

Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from https://www.laurence anthony.net/software.

Ädel, A. (2011). Rapport building in student group work. *Journal of Pragmatics*, *43*(12), 2932-2947.

Egbert, J., & Baker, P. (Eds.). (2019). *Using corpus methods to triangulate linguistic analysis*. London: Routledge.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *RANGE: A program for the analysis of vocabulary in texts*. Retrieved from http://www.vuw.ac.nz/ lals/staff/paul-nation/ nation.aspx.

Li, M., & Kim, D. (2016). One wiki, two groups: Dynamic interactions across ESL collaborative writing tasks. *Journal of Second Language Writing*, *31*, 25-42.

QSR International. (2018). *NVivo* (Version 12) [Computer software]. Melbourne, Australia: QSR International Pty Ltd. Retrieved from www.Qsrinternational.com.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.

**Casal, J. Elliott**

***The Pennsylvania State University***

**jec368@psu.edu**

Type of Contribution: Work-in-Progress Report

# Impacts of corpus-based English academic writing pedagogy on graduate students’ writing and decision-making

*Abstract*

In recent years, genre-based writing pedagogies have benefited from an increasing number of studies which report on corpus-driven pedagogical approaches to writing instruction (e.g., Chen & Flowerdew, 2018; Lee & Swales, 2006), particularly of academic English writing. Some scholars have adopted an integrated approach (e.g., Charles, 2007, 2011; Cortes, 2014; Dong & Lu, 2020) where corpus- and genre-based pedagogies represent “two equally important elements of class activity” (Charles, 2011, pp. 28-29). In most cases, studies on corpus-based writing pedagogies provide useful descriptions of the adopted pedagogical designs and present valuable student evaluations regarding the perceived benefits and difficulties of corpus-driven learning activity. In many cases these findings align with those of the larger tradition of corpus-based language instruction, that learners appreciate and perceive the benefits of corpus-driven activities even as they note that such activities are time consuming and often technically challenging. Overall, the results of these studies highlight the potential of such pedagogies to empower learners and raise awareness of discursive practices in spite of the considerable time commitment required, but few studies have reported on how corpus-driven pedagogies may impact student writers’ decision making processes or the texts that they produce.

The current work-in-progress report describes an ongoing study which continues this important pedagogical research tradition by analyzing the impacts of a corpus and genre analysis based pedagogical approach to writing

instruction on second language English doctoral students' decision-making when constructing disciplinary academic texts in English. The pedagogical intervention takes place in a course which integrates corpus- and genre-analysis activities (Charles, 2007, 2011) in a sixteen-week English for Academic Purposes (EAP) doctoral writing course at a large US university. Participants include thirteen doctoral student writers from ten disciplines (Adult Education, Architectural Engineering, Chemical Engineering, Ecology, Economics, Electrical Engineering, German Linguistics, Material Science, Physics, and Political Science) and six language backgrounds (Arabic, Chinese, Korean, Portuguese, Russian, and Spanish) recruited across two classes. Course materials were informed by a large-scale corpus analysis of the use of phrase-frames, shell nouns, reporting verbs, and complex noun phrases in the realization of writers' rhetorical aims in a rhetorical-move annotated corpus of 400 published research article introductions across four engineering and social sciences disciplines. Course materials were supplemented with discipline specific corpora constructed by students in course activity that involved both close analysis of expert writers' linguistic and rhetorical decisions and a variety of corpus-querying and analysis activities using AntConc (Anthony, 2019).

Data include student writing (a research article draft and later revision), two 18-question student surveys (5-point Likert) regarding their perceptions of and experiences with corpus and genre analysis based activities (mid-term and end-of- term), audio recorded class and group activities, and 45-minute individual semi-structured text-protocols targeting writers' linguistic and rhetorical decisions in the drafting and revision of their final text. In the present abstract a subset of the mid-term survey results are presented alongside a broad thematic analysis of the text protocol data. In future work the author will adopt a case-study approach to assess individual writers' development through comparative analysis of first and final drafts, survey responses, and analysis of learner participation in classroom and text-protocol activities for changes in intentionality and conceptual awareness of genre practices.

| Question | Mean | SD |
|---|---|---|
| 1. It was easy to build my own corpus | 3.83 | 1.11 |
| 2. It was useful to build my own corpus | 4.33 | 0.49 |
| 3. Analyzing moves in my corpus help improve my academic writing | 3.58 | 0.79 |
| 4. Analyzing my corpus with AntConc help improve my academic writing | 4.42 | 0.51 |
| 5. Using AntConc and move analysis together help improve my acad. writing | 4.33 | 0.49 |
| 6. Using AntConc and move analysis together help improve my acad. reading | 4.25 | 0.45 |
| 7. My instructor's assistance was important for learning to build and analyze my corpus | 4.92 | 0.29 |
| 8. I intend to use my corpus, AntConc, and move analysis in the future | 4.42 | 0.67 |

Table 1: Select learner perception survey questions and likert responses.

*Note*: Likert scale key: 1 = strongly disagree, 2 = disagree, 3 = neither disagree nor agree, 4 = agree, 5 = strongly agree.

Overall, the eight-question subset of the mid-term survey presented in Table 1 suggest that learners identified corpus activities as helpful in improving their writing (echoing previous findings) and useful in bridging the gap between linguistic forms and rhetorical goals in spite of the difficulties they reported in corpus compilation and rhetorical move analysis. Learners identified the role of the instructor as instrumental in such activity and report a strong interest in continuing to expand and work with their personal corpora, the AntConc tool, and move analysis on their own in the future. Students show a clearly higher rating for corpus-based activities over rhetorical analysis activities, which they largely explained by the way that move-step models aligned with their 'intuition.' These trends have strong connections to the researcher's observations made in analysis of classroom activity transcripts and text-protocols.

As in the survey, the students strongly indicated that engaging in rhetorical move analysis and corpus-based activities together made them more conscious of language form and rhetorical function in reading of academic texts and more intentional in their own writing, with many students commenting on their ability to usefully deploy their explicit knowledge of rhetorical and linguistic features in writing-based discussions with their academic advisors. While a small number of learners demonstrate reliance on imitation-like strategies in their own writing even after this pedagogical intervention, most show evidence of increasingly agentive adoption and integration of linguistic form-rhetorical construction conventions and developing conceptions of genre practices as social and situated, although to varying degrees. These issues will be investigated and reported more comprehensively in other contexts when data collection is complete. As a final note, several students also indicated extraordinary gratitude to Lawrence Anthony for the AntConc tool, which they viewed as a gateway to domain-specific language data to inform their decisions with highly specialized disciplinary writing questions.

### *References*

Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from https://www.laurence anthony.net/software.

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes, 6*(4), 289-302.

Charles, M. (2011). Using hands-on concordancing to teach rhetorical functions: Evaluation and implications for EAP writing classes. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning*, (pp. 26-43). London: Continuum.

Chen, M., & Flowerdew, J. (2018). Introducing data-driven learning to PhD students for research writing purposes: A territory-wide project in Hong Kong. *English for Specific Purposes, 50*, 97-112.

Cortes, V. (2011). Genre analysis in the academic writing class: with or without corpora? *Cuaderns de Filologia. Estudis Linguistcs, XVI*, 41-64.

Dong, J., & Lu, X. (2020). Promoting discipline-specific genre competence with corpus-based genre analysis activities. *Journal of English for Academic Purposes, 58,* 138-154.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*(1), 56-75.

**Čermáková, Anna**

***University of Cambridge/Charles University***

**anna.cermakova@ff.cuni.cz**

**Malá, Markéta**

***Charles University***

**marketa.mala@ff.cuni.cz**

Type of Contribution: Work-in-Progress Report

# Eyes and speech in English, Finnish and Czech children's literature

*Abstract*

Language of children’s literature has received surprisingly little attention (e.g. Stephens, 2004) and even less so cross-linguistically. One of the features that has been assumed and observed is a greater degree of explicitness than in the texts written for adult readers (Šebestová & Malá, 2019). Children, as readers are developing their reading and cognitive skills only gradually; Nikolajeva (2014) refers to them as ‘novice readers’, and the big, unexplored, question is how they make meaning out of texts. We can assume that communication between characters, both verbal and non-verbal, is key for meaning making. The ‘meaningfulness’ of fictional communication consists in “representing the *kind* of language which a reader can recognize, by observation, as being characteristic of a particular situation” (Leech & Short, 2007, p. 129). Here both verbal and non-verbal language play a role in allowing the reader, in our case the child reader, to imagine vividly and understand the situation.

This paper aims to explore the interplay between speech and its accompanying body language and it aims to do so across three languages - English, Finnish and Czech - using corpora that are comparable in terms of text-type (narrative fiction), medium (published books) and audience (children and teenagers). Body language is a wide and complex phenomenon, Korte (1997,

pp. 3-4) defines body language as “non-verbal behaviour […] which is ‘meaningful’”. We focus particularly on eye-behaviour. ‘Eye/eyes’ and their equivalents in Finnish and Czech, belong to the most frequently occurring body parts in fictional texts in general and fulfil a range of functions in supporting the speech and contributing to the characterization process. We explore instances where ‘eyes’ occur in the context of speech, and examine in what ways they interact with the speech, as in the following example, where the description of eye behaviour is in line with the emphatic reporting verb and the content of the following direct speech with an exclamation mark helping the reader to fully visualize the situation:

The **fires of fury** and **hatred** were **smouldering** in her small black **eyes**. ‘Matilda**!**’ she **barked**. ‘Stand up**!**’ Methodologically, we explore the lemmata EYE, SILMÄ and OKO and their collocates (rather than recurrent bundles). We are looking for their occurrences in the vicinity of speech (within a +/-5-word window). The three datasets we use are the following: for English we use the BNC subcomponent of children’s fiction (2 mil. words), for Finnish the subcomponent of the original Finnish texts in the Savokorpus (0.5 mil. words, provided by the courtesy of Prof. Anna Mauranen) and for Czech we use a subcorpus of Czech children’s books selected from the Czech National Corpus (2.8 mil. words, www.korpus.cz).

Preliminary results suggest that there are many similarities in terms of ‘eye language’ discourse functions across the languages: they are used with similar verbs (*look, open*), they are linked to expressions of emotion, both positive and negative, and light metaphors occur in all three languages.

### *References*

Korte, B. (1997). *Body language in literature.* Toronto: University of Toronto Press.

Leech, G., & Short, M. (2007). *Style in fiction.* Harlow: Longman.

Nikolajeva, M. (2014). *Reading for learning: Cognitive approaches to children’s literature.* Amsterdam: John Benjamins.

Šebestová, D., & Malá, M. (2019). Expressing time in English and Czech children’s literature: A contrastive n-gram based study of typologically distant languages. In J. Emonds, M. Janebová & L. Veselovská (Eds), *Language use and linguistic structure: Proceedings of the Olomouc linguistics colloquium 2018* (pp. 469-483). Olomouc: Palacký University.

Stephens, J. (2004). Linguistics and stylistics. In P. Hunt (Ed.), *International companion encyclopedia of children's literature* (pp. 99-111). London: Routledge.

**Coats, Steven**

***University of Oulu, Finland***

**steven.coats@oulu.fi**

Type of Contribution: Full Paper

## Dialect Corpora from YouTube

*Abstract*

Large corpora of geographically localized speech transcripts are an important resource for the analysis of regional variation in English (Szmrecsanyi, 2011), but few such corpora exist. This paper discusses the creation of dialect corpora from YouTube from automatic speech recognition (ASR) transcripts (Halpern et al., 2016), by using web scraping to identify YouTube channels of local government entities in North America and the British Isles. In addition, it introduces three new corpora created from government YouTube channels: one for the United States, one for Canada, and one for Great Britain and Ireland.

Coats (2019a) described a method for the creation of corpora from ASR transcripts of local government and community organization channels by sending multiple search terms to YouTube's API (Application Programming Interface), then downloading channel content using open-source tools. ASR transcripts can be used not only for frequency-based analyses of lexis and morphosyntax, but also for the investigation of speech timing phenomena (Coats, 2019b), as they include individual word timings.

Since early 2019, YouTube's API quotas have been reduced, making it less feasible to create a large corpus using this resource. Two alternative approaches for the identification of relevant channels can be used: First, pre-existing lists of local government websites (for example from the U.S. Census Bureau) can be scraped for links to YouTube channels. Second, lists of search terms can be sent directly to YouTube's public web interface (rather than the API). Both methods rely on the use of automated browser scripts. For the United States, a list of 35,924 websites was extracted from a comprehensive listing of 91,386 local government entities provided by the U.S. Census Bureau. These

websites, mostly homepages of cities, towns, or counties, were then scraped for links to YouTube channels. From the 2,376 channels identified in this manner, all English-language ASR transcripts were downloaded using YouTube-DL (Yen, Remite & Sergey, 2020) routed through the Tor network (see below; Loesing, Murdoch & Dingledine, 2010). Exact locations were assigned using a geocoder (Esmukov et al., 2018). The 335,017 transcript files resulted in a corpus of 1,236,298,290 words, corresponding to over 153,926 hours of video from locations in all 50 U.S. states and the District of Columbia. For Canada, a similar procedure was employed, using lists of municipalities or other local government entities retrieved from websites and databases of provincial or territorial governments. The procedure resulted in 40,966 transcript files from 443 channels in all 13 Canadian provinces and territories, with an aggregated word count of 123,499,887, corresponding to almost 15,321 hours of video. The U.S. and Canadian resources were combined with the corpus described in Coats (2019) to create a North American English corpus of just over 1.6 billion words. Figure 1 shows the locations of the channels from which transcripts were downloaded in this combined corpus.

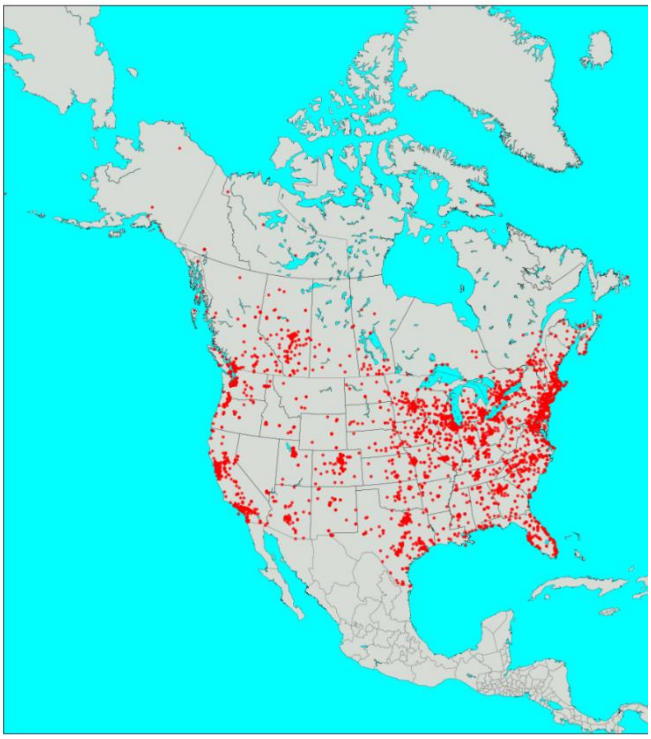


Figure 1: Locations of sampled channels.

For the British Isles, searches for council names were iteratively sent to YouTube’s search interface (e.g. “Dorset Council”, “East Ayrshire Council”, “Mayo County Council”, etc.), based on a list of the names of 413 local government authorities in England, Scotland, Wales, Northern Ireland, and the Republic of Ireland. After removal of duplicate results, automatically-generated channels, and false positives (e.g. the channel “Boston City Council” from the United States, rather than Lincolnshire), all transcripts of the 365 channel matches were downloaded, resulting in a corpus of 28,485 transcripts and 47,462,144 words, corresponding to 5,875 hours of video.

Web-scraping scripts must be carefully designed, tested, and refined in order to prevent false positives, and manual checking of channels is necessary to confirm that the content is from the targeted local government entity. For example, many municipal websites in North America are created using templates from commercial webpage-design software. If a municipality has no YouTube channel, but does not remove or alter the template used to create the municipal website, the page may include a link to the YouTube channel of the software provider.

The procedures described above require large numbers of HTTP requests to be sent to YouTube servers, which can result in the sender’s IP address being blocked for 24 or 48 hours or longer. To surmount this problem, scripts can be designed to send requests from multiple IP addresses, switching addresses after a certain number of calls. For users without access to multiple IP addresses and/or for whom the cost of acquiring multiple IPs via a virtual private network may be prohibitive, the Tor network can be used. Tor, an open-source software protocol for anonymous internet use, sends encrypted HTTP requests to a target via a randomized network of node servers. Periodically generating a new Tor connection changes the Tor “exit node” and thus the IP address of the server from which the request is passed YouTube. Using Tor reduces download speeds. To generate large corpora, several weeks or months may be necessary.

Although the methods detailed in this paper focus on the creation of speech corpora from specific locations, they may also be useful for the creation of other types of specialized corpora, for example pertaining to specified content, communicative situations, or speaker demographic attributes. In addition, the methods can be used to create corpora of video and audio files, which can then be subjected to various types of acoustic analysis.

### *References*

Coats, S. (2019a). A corpus of regional American language from YouTube. In C. Navarretta, M. Agirrezabal & B. Maegaard (Eds.), *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference* (pp. 79-91). CEUR.

Coats, S. (2019b). Articulation rate in American English in a corpus of YouTube videos. *Language and Speech*. Retrieved from: https://doi.org/10.1177/ 00238309 19894720

Esmukov, K. (2018). *Geopy* [Python module]. https://github.com/geopy/geopy

Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., & Bäuml, M. (2016). Contextual prediction models for speech recognition. In *Proceedings of INTERSPEECH 2017* (pp. 2338-2342).

Loesing, K., Murdoch, S. J., & Dingledine, R. (2010). A case study on measuring statistical data in the Tor anonymity network. In R. Sion, R. Curtmola, S. Dietrich, A. Kiayias, J.M. Miret, K. Sako, & F. Sebé (Eds.), *Financial Cryptography and Data Security: FC 2010 Workshops, RLCPS, WECSR, and WLC 2010 Tenerife, Canary Islands, Spain, January 2010, Revised Selected Papers* (pp. 203-215). Berlin: Springer.

Szmrecsanyi, B. (2011). Corpus-based dialectometry: A methodological sketch. *Corpora, 6*(1), 45-76.

Yen, C. H., Remite, A., & Sergey M. (2019). *Youtube-dl* [Software]. Retrieved from: https://github.com/rg3/youtube-dl/blob/master/README.md.

**Collins, Luke**

***Lancaster University, UK***

**l.collins3@lancaster.ac.uk**

**Semino, Elena**

***Lancaster University, UK***

**e.semino@lancaster.ac.uk**

**Demjén, Zsófia**

***University College London, UK***

**z.demjen@ucl.ac.uk**

Type of Contribution: Full Paper

## Corpus linguistics and clinical psychology

Examining the psychosis continuum

*Abstract*

We present our work applying methods from corpus linguistics to the field of clinical psychology. Our study focuses on reports of voice-hearing: the experience of hearing voices that others cannot hear. We investigate contrasting experiences of those who find spiritual meaning in their voice-hearing and those who find their experiences distressing to the point that they seek clinical support. We contextualize our work among debates in psychology around the notion of the 'psychosis continuum', applying keyness analyses to identify areas of similarity between two groups of voice-hearers. Furthermore, in looking at the dispersion of features at the individual level we investigate how we can extend this discussion of 'similarity' to consider 'continuity'.

We worked with the 'Hearing the Voice' team at Durham University in studying semi-structured interview data with 67 voice-hearers. In analyzing the language choices the voice-hearers made in reporting their experiences, we

demonstrate how corpus methods can contribute to a broader investigation of psychosis. The experience of hearing voices - or Auditory Verbal Hallucinations (AVHs) as they are labelled in clinical psychology - is primarily associated with schizophrenia and other mental health difficulties (Van Os & Reininghaus, 2016), however AVHs also occur as a positive and meaningful experience for many who do not seek or require clinical care, such as spiritualists. As such, our interview data represents the experiences of 27 self-identified spiritualists and 40 voice-hearers who are registered with the clinical services provided by NHS England.

The ‘psychosis continuum’ model posits that the broad experiences of clinical (i.e. service users) and non-clinical (e.g. spiritualists) voice-hearers can be conceptualized as a continuum and that the experiences of such groups vary in terms of, for example, distress, vividness and duration (Waters & Fernyhough, 2019). Research from psychology has reported differences between clinical and non-clinical groups in terms of how voice-hearers interpret the experience and the degree of control the voice-hearer feels they have over their AVHs (Baumeister et al., 2017), which has implications for the type of clinical care they receive. However, such research has relied on psychometric measures to identify differences, which are reported in terms of statistical significance and what is reported as ‘similarity’ is, more accurately, the absence of a statistically significant difference. Our approach sets out to develop measures of degrees of ‘similarity’ and ‘difference’ and to consider the extent to which aspects of clinical and non-clinical experiences can be said to be 'continuous'.

We carried out keyness analyses of the spiritualist and service user data at the level of semantic domains, using the USAS tagger (Rayson, 2008). We performed a direct comparison between the groups, as well as a comparison of each cohort with a third reference corpus (a corpus of oral history interviews taken from the BNC1994 (Aston & Burnard, 1998)) in order to identify key semantic domains that were overused by one or both of the participant groups. This provided an initial indication of similarity and difference at the group level. Focusing on semantic domains that corresponded with aspects of voice-hearing that had been previously shown to be pertinent to the ‘psychosis continuum’ - such as perceived level of control, level of distress etc. - we looked at the dispersion of terms within those key semantic domains at the individual level to investigate ‘continuity’ across the clinical and non-clinical groups.

Our results showed that in cases where there were group-level difference in the frequency of terms relating to prominent themes between service users and spiritualists, there was still overlap in values for individuals, suggesting continuity between the clinical (service users) and non-clinical (spiritualists). Furthermore, applying simple statistical models, we found ‘discontinuity’ between participants within the same clinical group i.e. between a smaller group of clinical voice-hearers, rather than between clinical and non-clinical voice-hearers. This highlighted that certain aspects of the experience, such as ‘negative affect’, were of particular relevance to specific participants, which could be used to direct the clinical care they receive.

As members of the ESRC Centre for Corpus Approaches to Social Science, our collaboration with the 'Hearing the Voice' team reflects our commitment to working with interdisciplinary teams to investigate communication about health and illness. The process has emphasized the importance of working with experts in the field, who are familiar with the leading research and can support corpus linguistics in making a meaningful contribution. There are, however, some limitations on the extent to which we can claim to be introducing innovation to the field given that we first have to demonstrate the applicability of the corpus method to the existing literature. Nevertheless, interdisciplinarity contributes to the development of research in multiple areas and offers the opportunity to extend the application of corpus methods to a new area of research and contribute to debates in other fields.

## *References*

Aston, G. & Bernard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baumeister, D., Sedgwick, O., Howes, O., & Peters, E. (2017). Auditory verbal hallucinations and continuum models of psychosis: A systematic review of the health voice-hearer literature. *Clinical Psychology Review, 51*, 125-141.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519-549.

Van Os, J., & Reininghaus, U. (2016). Psychosis as a transdiagnostic and extended phenotype in the general population. *World Psychiatry, 15*(2), 118-124.

Waters, F., & Fernyhough, C. (2019). Auditory hallucinations: Does a continuum of severity entail continuity in mechanism? *Schizophrenia Bulletin, 45*(4), 717-719.

**Dong, Min**

***Beihang University***

**mdong@buaa.edu.cn**

**Fang, Alex Chengyu**

***City University of Hong Kong***

**acfang@cityu.edu.hk**

Type of Contribution: Full Paper

## Shell nouns as register-specific discourse devices

*Abstract*

Shell nouns (SNs), a special class of abstract nouns such as *fact, problem* and *suggestion*, have been extensively studied because of their referential and stance-taking functions in discourse. The present article focuses on SN+*that* constructions that perform an intra-clause cataphoric function with a specific view to their deployment in spoken and written discourse, aiming to provide an empirically verifiable account of their uses across speech and writing (Yamasaki, 2008; Benítez-Castro, 2014; Flowerdew & Forest, 2015, p. 101; Schmid, 2000, p. 380) through a publicly available corpus comprising both written and spoken texts of varying registers. It intends to address three research questions:

(1) What is the distribution characteristic of SN+*that* constructions in a range of varying settings in speech?

(2) How does the distribution characteristic of SN+*that* constructions in speech compared with that in writing?

(3) Is there any discernible principle governing the contextual deployment feature of SN+*that* constructions in a spectrum of varying spoken and written registers?

The current study opted for the International Corpus of English (ICE), which is a corpus-based linguistic project that aims to investigate the grammatical

differences and similarities across a set of national and regional varieties of English used as either a first majority language or an official language (Greenbaum, 1996). While several component corpora have been constructed, the British component (ICE-GB) is the only one that has been grammatically tagged for part-of-speech information and then syntactically parsed (Nelson, Wallis & Aarts, 2002). The overall design is balanced between spoken and written modes of production, text categories, allocated number of samples and standard length of texts to facilitate comparisons and contrasts. During the syntactic annotation of ICE-GB, each parsed tree was facilitated by computer software and manually validated, hence enjoying a high degree of consistency as well as accuracy. In our study, the SN+*that* constructions in ICE-GB were extracted from the corpus according to the syntactic annotation NPPO-CL (sub) before manual checking was applied to ensure reliability of analysis.

First of all, our investigation in the corpus distribution of SN tokens and types has shown that speech has a lower use of SNs than writing in terms of normalized occurrences per 100,000 word tokens although the difference is only marginal in statistical terms. The fact that speech and writing are observed in both polarities with higher and lower SN tokens, serves as strong indication that SN+*that* constructions are not a function of production modes in terms of speech and writing but one of a registerial setting, contrary to past suggestions.

Within the spoken register, it is found that the occurrences of SN+*that* construction increase from private dialogues to public dialogues, and from unscripted to fully scripted monologues. Noticeably, private dialogues exhibit a much lower number of SN tokens than the total average of speech.

The same observations apply to the written data: the printed section demonstrates a higher SN occurrence than that of the non-printed section, along with the increase of specificity of subject matter and the increase in argumentativeness. The change is similarly conditioned by degrees of preparedness of subject matter: time-constrained productions yield a lower use of SNs than unconstrained productions. Within the printed, a pattern of increase can be observed across the six sub-categories from descriptive registers to persuasive registers. This finding may possibly point back to the suggestion that SNs are discursive devices preferred in argumentative discourse. Additional evidence for this possibility comes from the observation that persuasive writing displays the highest use of SNs, suggesting a special preference in argumentative discourse for SNs as a meta-discursive device. The same logic seems to suggest that reportage is also part of argumentative discourse, explainable through the discrete use of SNs as stance indications.

The above observations reveal that SNs are found to be equally likely to occur in speech and writing. More specifically, SNs are preferred in registers contextually conditioned by more formal participant role relationship, longer preparation time for subject matter, more specialized subject matter and more persuasive goal orientation, therefore a register-specific device performing meta-discursive functions in productions discursive in nature, irrespective of mode. Conversely, the empirical evidence suggests that SNs are a registerial indicator; a formal register tends to be accompanied by a high occurrence of SN+*that* constructions.

However, a higher occurrence of SNs does not incur a higher lexical diversity in terms of types, contradicting the usual assumption that more formal registers are accompanied by higher lexical diversity of SNs. This phenomenon reinforces the suggestion that the SN+*that* construction is less an issue of lexis but perhaps pertains more to the grammatical system. Our study has identified a common core set of SNs which accounts for a majority of the uses across speech and writing, leading to the suggestion that SN+*that* should perhaps be regarded as a ubiquitous language phenomenon. Furthermore, the Jaccard Similarity Index (JSI) score effectively means that across the two sets of SN types of similar size, just about one third (36.7%) are identical, suggesting that the two mediums of discourse have significantly different preferences in the choice of SN types. In addition, between speech and writing, thus, we observe that less than half of the SN types are uniquely used in either mode. Etymological differences between the two modes speculated in past studies are not found to be evidently significant.

The findings arising from the current study have also called for further investigations. For one thing, we have focused on the structural and lexical properties of SNs as a register-specific discourse device. A natural extension would be to investigate SN+*that* constructions from the perspective of grammatical metaphor defined in Hallidayan terms and find out how SNs as instances of grammatical metaphor influence different cognitive construals across varying spoken and written registers with regard to their semantic typing. This is where the linguistic analysis of SNs as a prominent phenomenon of English can be expected to yield understanding about a 'meta-linguistic' model that will insightfully explain the formation and the contextual configuration of register.

## *References*

Benítez-Castro, M. Á. (2014). Formal, syntactic, semantic and textual features of English shell nouns: A manual corpus-driven approach. In A.A. Sintes & S.V. Hernández (Eds.), *Diachrony and synchrony in English corpus linguistics* (pp. 171-203). New York: Peter Lang.

Flowerdew, J., & Forest, R.W. (2015). *Signaling nouns in English: A corpus-based discourse approach*. Cambridge: Cambridge University Press.

Greenbaum, S. (1996). *Comparing English worldwide: The International Corpus of English*. Oxford: Oxford University Press.

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam: John Benjamins.

Schmid, H. J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin: Mouton de Gruyter.

Yamasaki, N. (2008). Collocations and colligations associated with discourse functions of unspecific anaphoric nouns. *International Journal of Corpus Linguistics*, *13*(1), 75-98.

**Dykes, Natalie**

***Friedrich-Alexander-Universität Erlangen-Nürnberg***

**natalie.mary.dykes@fau.de**

**Heinrich, Philipp**

***Friedrich-Alexander-Universität Erlangen-Nürnberg***

**philipp.heinrich@fau.de**

**Blombach, Andreas**

***Friedrich-Alexander-Universität Erlangen-Nürnberg***

**andreas.blombach@fau.de**

Type of Contribution: Work-in-Progress Report

## Independent argumentation schemes?

Transferring argument queries from Brexit to environment tweets

*Abstract*

We present a corpus-based study of argumentation on Twitter. The large-scale analysis of argumentation patterns has enjoyed increasing popularity in recent years, with most projects relying heavily on machine learning. Corpus linguistics so far has played a relatively minor role; the main focus of attention usually lies on the development of logical representation frameworks and of identifying the argumentative function of utterances (e.g. *premise, conclusion*) and their relations (*attack*, *support*, etc.; cf. Lippi & Torroni, 2016). However, these approaches usually work best on clearly structured texts and "well-behaved" arguments which follow a normatively coherent structure. An argument of this kind consists of one or several premises and a conclusion which is always true given the truth of the premise(s). With the increasing prominence of social media, the field has shifted its attention towards considerably more noisy data,

usually resulting in very low precision scores for the typical extraction and mapping tasks (Goudas et al., 2014).

The first reason that argumentation on platforms like Twitter is difficult to study is related to their medial characteristics: the texts are very short and typically contain high amounts of non-standard language, which also affects standard processing and tagging tasks. Secondly, the argumentative make-up of everyday utterances does not follow the strict normative scheme: essential reasoning steps are typically left implicit; and the conclusion is not strictly implied by the premise(s) (cf. Walton et al., 2008). A classic example is arguers attempting to strengthen their position by discrediting the opponent (*ad hominem*).

In a previous effort, we developed corpus queries to capture arguments in English tweets about the 2016 Brexit referendum and the later developments as reflected by Brexit tweets from 2019 (Dykes et al., 2019; Dykes et al., forthcoming). We build on the IMS Open Corpus Workbench (CWB, Evert & Hardie, 2011). CWB allows its users to define macros and word lists, which can be re-used in different queries. We developed a Python-based wrapper[1] to deal with anchored CQP queries as well as a web-based application where users can inspect concordances, refine queries and obtain frequency tables for selected tokens or token sequences.

In current work, we explore a new corpus of 515,861 English tweets (17,263,183 tokens) containing the search string “environment”, collected via Twitter’s Streaming API (April to September 2019). In order to restrict the corpus to tweets that are actually about the environment, we excluded tweets where *environment* was preceded by determiners except *the* and *our*, pronouns other than *our*, and a number of content words and hashtags indicating that the word was used in a non-ecological sense (*built, designed, hostile, urban, work, #IT, today’s* etc.). The amount of near duplicates is relatively low (around 10%) and a stable daily rate of around 5000 tweets (see Schäfer et al., 2017 for details on our deduplication algorithm). One significant peak can be observed on June 5 (“World Environment Day”), with more than 40,000 tweets.

[1] https://pypi.org/project/cwb-ccc/

While the original queries were developed on the Brexit corpus, they were designed to capture everyday argumentation independently of a particular topic. Instead, they are built to represent a specific type of linguistic realization for a logical relation between entities and concepts (*X causes Y*, *X and Y are similar*, *all members of group X have property Y* etc.). All of these patterns can be assumed to be general reasoning steps as typical components of everyday argumentation. We therefore hypothesize that our queries will be applicable to thematically diverse datasets, given that the text genre, linguistic structure and register are kept relatively similar. Our present aim is to test this assumption on the environment tweets.

Queries contain regular lexico-grammatical patterns like POS and wordlists representing lexico-semantic categories. The wordlists are the only element that is somewhat topic-specific, in that they contain lemmas commonly found in the Brexit corpus. Thus, they may need expansion and modification for other corpora. The query below yields instances of *ad hominem*; an argumentative strategy that can be expected to occur very commonly in tweets about controversial issues, with the aim of discrediting the opponent’s perspective by insulting them as a person:

```
<np>@0[::][]*[$person|pos="P|Z"]@1:[]*</np>
[lemma = “““be”] “““too” [lemma = $stupid] “““to”
<vp>@2:[]@3:[]*</vp>
```

This query matches utterances of the type “[person] is too [negative attribute] to VP”, which is a very straightforward way to express the scheme. $person is a word list containing nouns referencing people, while $stupid consists of adjectives like *stupid, dumb, daft…* Interesting query parts are marked by anchor points; indicated by @0 etc. These anchors allow for focusing words or sequences to obtain frequency information. They also enable further processing, e.g. for logical analysis, where one would want to isolate particular roles and relations. For instance, the sequence between tokens @0 and @1 consists of the entire NP which corresponds to the entity being attacked while the second sequence matches the VP - the action they are allegedly intellectually incapable of performing. On the Brexit corpus, our pattern matches results of the following type:

(1) donald trump is too stupid to know the significance of Brexit

Examples from the environment data include:

(1) y'all too scared to go against that so y'all just attacking Indigenous efforts🖕

(2) your generation is too ignorant to let CHANGE happen!

The environment corpus is considerably smaller than our Brexit data. As a result, highly specialized queries that are rare in the Brexit corpus yield extremely few or no hits on the new dataset. This is not necessarily due to limited transferability of our queries, but likely at least partially a consequence of our general aim for high precision while somewhat neglecting recall. Figure 1 below shows an overview of argument queries in a relative comparison between the corpora. While we currently have around 70 queries in total, the figure only includes those with a minimum frequency of 0.2 per 1,000 tweets.

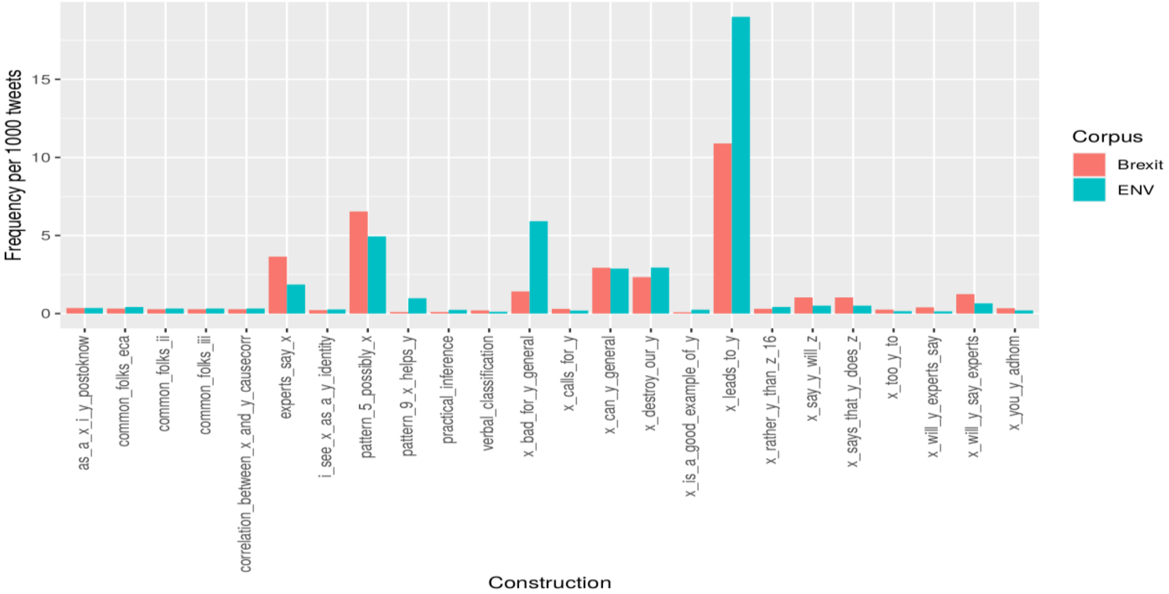


Figure 1: Top argument patterns per 1000 tweets as identified by our queries.

While the frequency of most specialized patterns is expectedly very low for both datasets, more general patterns tentatively suggest the presence of potential systematic differences in broad argumentative strategies. In the environment data, the causal pattern of the type *X leads to Y* was found to be considerably more common, as was the case for *X (is) bad for Y* and *X destroys our Y*. All three of these queries typically realize some form of arguments from bad consequences; arguing that an action ought to be avoided because of its anticipated negative results. For Brexit tweets, the more common patterns tend to be related to citing actors (*X says that Y does Z, X will Y say experts, experts say X*) and queries relating to modality; in particular potentiality (*possibly X, X can Y*).

Currently, our progress only allows for very broad conclusions: it appears that the status of authorities plays a more important role for Brexit tweets; especially in the context of high-prestige (or at least highly prominent) actors being cited. The prominence of modal expressions modifying a proposition is in line with our findings in a closer analysis of the Brexit data which has also found uncertainty to play a major role (Dykes et al., forthcoming). In contrast, the higher prominence of explicit causality indicates that tweets on the environment might construe a higher degree of confidence in the expected outcomes of undesired actions.

### *References*

Dykes, N., Heinrich, P., & Evert, S. (2019). Arguing Brexit on Twitter: A corpus linguistic study. *European Conference on Argumentation 2019*, Groningen, Netherlands.

Dykes, N., Heinrich, P., & Evert, S. (forthcoming). Labour are red, Tories are blue, some arguments stay, and our methods are new: Reconstructing Twitter arguments with corpus linguistics. In *Proceedings of ICAME40*.

Evert, S., & Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of CL*.

Goudas, T., Louizos, C., Petasis, G., & Karkaletsis, V. (2014). Argument extraction from news, blogs, and social media. In *SETN 2014*.

Lippi, M., & Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, *16*(2), 1-25.

Schäfer, F., Evert, S., & Heinrich, P. (2017). Japan’s 2014 general election: Political bots, right-wing internet activism and PM Abe Shinzō’s hidden nationalist agenda. *Big Data, 5*(4), 294-309.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge: Cambridge University Press.

**Ebeling, Jarle**

**University of Oslo**

***jarle.ebeling@usit.uio.no***

Type of Contribution: Work-in-Progress Report

## Think it. Say it. Reported.

A century of reporting verbs in British fiction (1900-2019)

*Abstract*

This work-in-progress report investigates reporting verbs in 20th and 21st century British fiction. The investigation is limited to reporting verbs in direct speech (DS) and thought (DT), e.g. *cried* in (1) and *thought* in (2), respectively.

(1) ““Give us another call!” he cried.

(2) ““Why not?’ he thought.

In his book on Dickens and the Suspended Quotation, Lambert (1981) observes that the use of *cry* as a reporting verb has been, and still is, dropping in use. Based on his observation I wanted to find out whether this decrease affects only some reporting verbs, and whether some verbs become more frequent as others decrease in use.

And, if we can detect a decrease in the use of certain reporting verbs, is this tendency equally strong regardless of who is reported speaking. This last question latches on to something Underwood (2019, p.123) and Underwood et *al*. (2018) address, namely whether we can observe a “growing blurriness of gender boundaries” in fiction?

With these research questions as my point of departure, reporting verb and speaker (*he* or *she*) were extracted from passages of dialogue in the Corpus of British Fiction (CBF). The CBF contains 882 novels of general fiction, by 411 different authors, and approx. 70 million words. It is tagged with the CLAWS part-of-speech tagger (Garside & Smith, 1997) and lemmatized with

TreeTagger (Schmid, 1994). Table 1 gives an overview of number of texts/words per decade.

| Decade of publication | Number of texts | Number of words |
|---|---|---|
| **1900-1909** | 80 | 7,444,527 |
| **1910-1919** | 93 | 7,465,198 |
| **1920-1929** | 109 | 9,386,574 |
| **1930-1939** | 93 | 7,817,215 |
| **1940-1949** | 50 | 4,436,032 |
| **1950-1959** | 55 | 3,781,486 |
| **1960-1969** | 59 | 4,500,186 |
| **1970-1979** | 54 | 4,060,160 |
| **1980-1989** | 83 | 4,940,756 |
| **1990-1999** | 88 | 5,479,874 |
| **2000-2009** | 54 | 5,319,611 |
| **2010-2019** | 64 | 6,363,988 |
| **Total** | 882 | 70,995,607 |

Table 1: The number of texts and words per decade.

494,000 reporting verbs were automatically extracted from the corpus. The speaker is *he* in 26% of the cases, *she* in 19% and other/unknown in the remaining cases. The unknown category is typically made up of other pronouns and proper nouns.

Figure 1 traces the amount of reporting verbs in the 120-year period covered by the CBF. The numbers have been normalized by dividing the frequency of verbs per decade by the total number of both direct and free direct speech and thought in that decade.

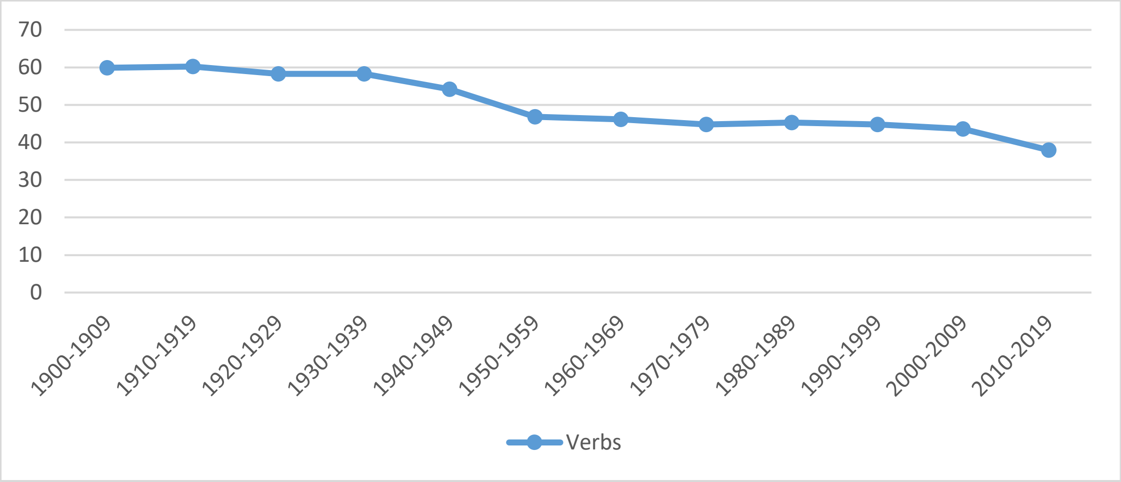


Figure 1: Reporting verbs per decade.

Manual scrutiny of 5,000 randomly selected instances of both direct and free direct speech and thought showed that a little more than half of all occurrences (57%) are free direct speech and thought, as in "*Oh, for God's sake.*"

The figure shows a steady decrease in the use of reporting verbs from 1900 to the present. There may be several reasons for this, e.g. the structure of the corpus or the ways in which direct speech is marked up. A manual inspection of 2,000 randomly selected positives and negatives revealed 93% precision and 74% recall.

As has been noted by several scholars, e.g. de Hahn (1996), *say* is the reporting verb par excellence, and, indeed, it accounts for more than 60% of all occurrences of reporting verbs in the CBF (300,000 of the 494,000). The second-most frequent verb is *ask* which accounts for 5.5% of the occurrences. *Whisper* at rank ten accounts for 0.9% and *laugh* at rank thirty 0.3%.

Reporting verbs sorted by frequency: *say* (60.7%)*, ask* (5.5%)*, reply, cry, answer, tell, add, exclaim, murmur, whisper* (0.9%)*, remark, go on, continue, think, repeat, agree, begin, shout, explain, mutter, demand, enquire, call, suggest, observe, declare, admit, protest, announce, laugh* (0.3%)

The infrequency of verbs ranked below the top ten in the later decades means that we get few instances per decade to draw our conclusions on, and this number becomes even lower when we only look at the ones where *he* or *she* is the speaker.

To investigate Lambert’s claim about the decreasing use of *cry* as a reporting verb, the data were split into three periods, 1900-1939, 1940-1979 and 1980-2019, and sorted by raw frequency.

1900-1939
*say, ask, reply,* ***cry****, answer, exclaim, remark, murmur, think, whisper, add, continue, repeat, go on, enquire, begin, agree, demand, explain, mutter, tell, suggest, observe, declare, laugh, shout, admit, call, retort, return*

1940-1979
*say, ask, tell, reply, add, answer,* ***cry****, exclaim, agree, go on, murmur, think, explain, shout, whisper, repeat, call, begin, remark, demand, suggest, enquire, continue, mutter, announce, protest, snap, smile, admit, sigh*

1980-2019
*say, ask, tell, reply, add, whisper, shout, continue, murmur, mutter, call, go on, begin, agree, snap, explain,* ***cry****, repeat, demand, answer, suggest, exclaim, announce, enquire, yell, protest, admit, comment, remark, observe*

A quick perusal of the lists shows that there is remarkable stability in the use of reporting verbs in the period covered. In fact, several scholars (e.g. Kytö et al., 2006; Busse, forthcoming) make this point and argue that we should not disregard stability as a powerful factor in the history of language in our (eager) pursuit of change. We can notice some changes, though, not least in relation to *cry*. *Cry* has dropped in popularity, while another verb with overlapping meaning, *shout*, has gained in popularity. Is it the case that *shout* is about to oust *cry* as a way of expressing a strong emotion in British fiction?

The stability is also apparent when ranking the verbs according to speaker, *she* or *he*, for the whole period (1900-2019).

Female speaker (*she*)
*say, ask, reply, cry, answer, tell, murmur, add, whisper, think, exclaim, agree, repeat, remark, go on, begin, continue, explain, shout, enquire, mutter, call, suggest, demand, observe,* ***protest****, admit, snap, declare,* ***smile***

Male speaker (*he*)

*say, ask, reply, cry, answer, tell, exclaim, add, murmur, remark, go on, whisper, continue, repeat, agree, begin, think, shout, mutter, explain, demand, enquire, call, suggest,* ***observe****, declare,* ***announce****, admit, laugh, snap*

If we compare the two lists, only *protest* and *smile*, which occur among the top 30verbs with *she*, do not occur among the top 30 with *he*. Similarly, two verbs used by *he* do not make it to the top 30 list used by *she*: *observe* and *announce.* More interesting perhaps is the rank of *think*. If we take these two lists at face value, *she* seems to be thinking more than *he*, or, more accurately, authors depict women through direct thought more than they do men.

Regarding the potentially ““growing blurriness of gender boundaries”, this is difficult to establish based on the CBF due to the drop in overall frequency of verbs other than *say*. Figure 2, for instance, showing the development of the use of *think,* seems to indicate this blurriness of gender boundaries, but as there are few occurrences of *he/she thought* in the later decades, more data is needed to establish blurriness of gender with more certainty.

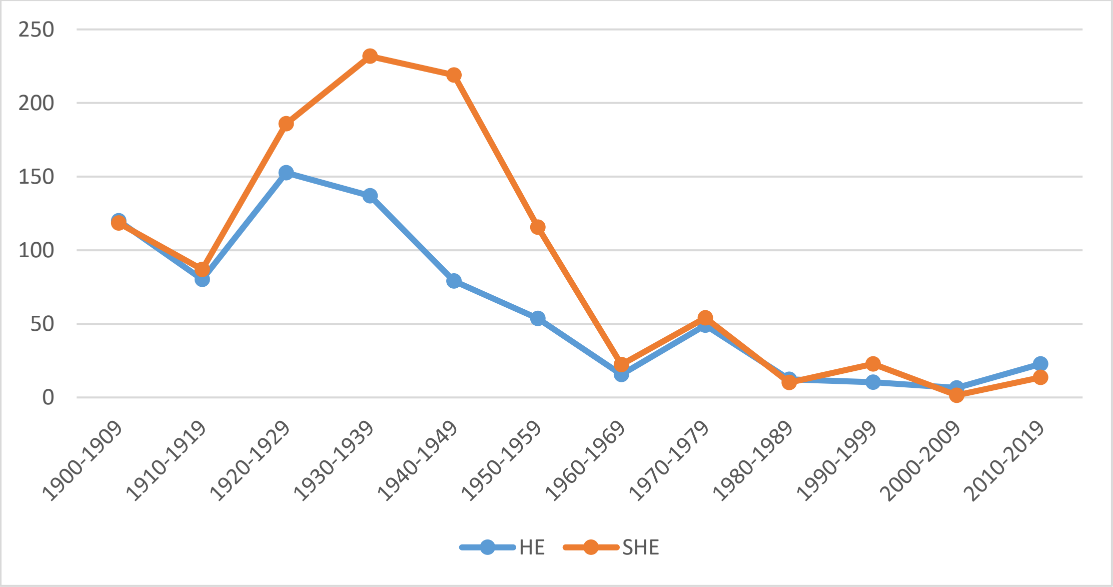


Figure 2: Normalized frequency of *he / she* THINK.

What Figure 2 does show, however, is that the use of direct thought was more often a characteristic of females than males between 1920-1960, but seemingly not before or after that period.

- To sum up, direct speech and thought seem to be decreasing in use, perhaps at the expense of free direct speech and thought;
- *say* is more frequently used as a reporting verb than all the other verbs put together;
- there is great stability in the use of reporting verbs in the period covered, and also when it comes to who is speaking, i.e. either *he* or *she;*
- the sharp drop in use of individual verbs overall, e.g. *think*, makes it difficult to make definite claims about any blurriness of gender boundaries, although there is some evidence that the gap between the genders is narrowing when we move closer to the present.

## *References*

Busse, B. (Forthcoming). *Speech, writing, and thought presentation in 19th-century narrative fiction: A corpus-assisted approach*. Manuscript. Oxford University Press.

de Haan, P. (1996). More on the language of dialogue in fiction. *ICAME Journal 20*, 23-40.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102-121). Longman.

Kytö, M., Rydén M., & Smitterberg. E. (2006). Introduction: Exploring nineteenth-century English - past and present perspectives. In M. Kytö, M. Rydén, & E. Smitterberg (Eds.), *Nineteenth-century English: Stability and change* (pp. 1-16). Cambridge: Cambridge University Press.

Lambert, M. (1981). *Dickens and the suspended quotation*. Yale University Press.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK. Retrieved from:

https://www.cis.uni-muenchen.de/~schmid/tools/ TreeTagger/data/tree-tagger1.pdf

Underwood, T. (2019). *Distant horizons: Digital evidence and literary change*. Chicago: The University of Chicago Press.

Underwood, T., Bamman, D., & Lee, S. (2018). The transformation of gender in English-language fiction. *Cultural Analytics Feb. 13*.

**Egan, Thomas**

***Inland Norway University of Applied Sciences***

**thomas.egan@inn.no**

# Telling in English and Norwegian

Type of Contribution: Work-in-Progress Report

*Abstract*

This paper presents the results of a study of ditransitive and prepositional dative constructions containing the cognate verbs English *tell* and Norwegian *fortelle*, using data from the English-Norwegian Parallel Corpus (ENPC: see Johansson, 2007, p. 10). It is part of a larger study of a handful of cognate verbs that display the dative alternation in both languages, coding actions of giving, sending, bringing, lending and serving, as well as telling. An analysis of GIVE constructions in the two languages shows that these are remarkably similar, both in their semantics and their distribution (Egan, forthcoming).

The reason for selecting cognate verbs for the study is grounded in the assumption that translators, in addition to attempting to render the semantic and pragmatic import of their source texts, will tend to employ congruent syntactic constructions where these are available in the target language (see Ebeling, 1998, 169). Given this premise, an analysis of the translation correspondences of the TELL constructions may be expected to throw further light on the similarities and differences between their distributions in the two languages. The TELL verbs in the two languages differ from the GIVE verbs analyzed in Egan (forthcoming) in two main respects. In the first place, they are less likely to occur in prepositional object constructions, at least in English. In the second place, as pointed out by Mukherjee (2004: 127), the direct object is much more likely to take the form of a clause, as in (1) and (2).

(1) He *did* once *tell* me that he hated shaking hands. (RDA1)
Han *fortalte* meg en gang at han hatet å håndhilse. (RDA1T)

(2) One time he *told* me what the name of the town meant. (NG1)
En gang *fortalte* han meg hva navnet på byen betydde. (NG1T)

The constructions in the translations in (1) and (2) mirror those of the originals. Preliminary results show that there is 70% mutual correspondence (see Altenberg, 1999) between the two verbs in constructions containing nominal direct objects, but just 45% and 60% in constructions with finite direct objects like those in (1) and (2), respectively. There is very little correspondence in constructions with objects in the form of direct speech, and none whatsoever in the case of non-finite clausal objects, which only occur with *tell.*

## *References*

Altenberg, B. (1999). Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (249-268). Amsterdam: Rodopi.

Ebeling, J. (1998). Using translations to explore construction meaning in English and Norwegian. In S. Johansson & S. Oksefjell (Eds.), *Corpora and cross-linguistic research: Theory, method and case studies* (169-195). Amsterdam: Rodopi.

Egan, T. (forthcoming). Giving in English and Norwegian: A contrastive perspective. In M. Röthlisberger, E. Zehentner & T. Colleman (Eds.), *Ditransitive constructions in Germanic languages*. Amsterdam: John Benjamins.

Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Mukherjee, J. (2005). *English ditransitive verbs: Aspects of theory, description and a usage-based model.* Amsterdam: Rodopi.

**Fuchs, Robert**

***University of Hamburg***

**robert.fuchs@uni-hamburg.de**

**Rautionaho, Paula**

***University of Eastern Finland***

**paula.rautionaho@uef.fi**

Type of Contribution: Full Paper

# Dynamic uses of stative verbs

## Recent change in spoken British English

*Abstract*

The recent diachronic increase in the frequency of the progressive has been attributed to the spread from its prototypical domain, dynamic verbs, to new domains, such as stative verbs (e.g. Mair, 2006). However, an analysis of their meaning in context reveals that we can distinguish stative uses of such ‘stative’ verbs as well as dynamic uses (e.g. *they are hurting* - ‘they are in pain’ vs. *they are hurting the animal* - ‘causing the animal pain’). This distinction, although well known, has so far rarely been taken into account in research on stative progressives in English, or at least no clear distinction is made. For instance, Leech et al. (2009, p. 129 fn.) point out that stative verbs ““allow both stative and dynamic interpretations”, but it remains unclear whether that distinction was actually made in the analysis of the data.

For stative progressives, there is real-time evidence of increasing use in 19th-century British English (e.g. Smitterberg, 2005), but 20th-century data shows that this increase may be halting (Leech et al., 2009). Extending and refining such analyses, our own study of stative uses of stative progressives indicates that they have not generally become more frequent in spoken British English (BrE) between the 1990s and the 2010s, but that particular verb lemmata (e.g. EXPECT, THINK) are indeed used more frequently in the

progressive, which we argue is due to their increasing use as politeness/hedging markers (Rautionaho & Fuchs, 2020).

In the present study, we extend this work to dynamic uses of stative progressives and ask whether they have become more frequent from the early 1990s to the early 2010s, based on the spoken, demographic sections of the old and new British National Corpus (Love et al., 2017). Our research questions are as follows: (i) have dynamic uses of stative progressives become more frequent in the present day, compared to the early 1990s, (ii) has there been any change in the collostructional preferences of dynamic uses of stative verbs, and (iii) is the impression of increased use of stative progressives due to increase in frequency of dynamic uses of stative progressives. Since the use of stative progressives per se has not increased in the past 20 years, we now focus on dynamic uses of such progressives to see whether trends in their usage have led to the impression of increase.

For details on data extraction, see Rautionaho and Fuchs (2020), as the present study is based on the residuals of the earlier study. Importantly, we restricted the analysis to a variable context where a progressive could potentially occur (excluding, for example, imperatives, constructions with modal verbs, and idiomatic expressions such as *you know*), and separated dynamic uses from stative uses; if the reading of the token in its context entailed any energy involved, the token was categorized as dynamic. Thus, instances such as HANG in the sense of ‘to fasten’ and HAVE in senses other than ‘possess’ were excluded from the dataset used for Rautionaho and Fuchs (2020), but make up the dataset for the present study. In total, the present dataset contains 1,289 dynamic uses of stative verbs (both progressive and simple).

The frequency results indicate that, overall, there is a small statistically significant increase between the proportion of progressive usage of dynamic stative verbs in the two corpora (BNC1994DS 28.5% vs. BNC2014S 31.6%; LL=5.80, p<0.05). We find 17 lemmata that allow dynamic reading and occur in the progressive and simple at least once; we also find considerable variation between the two corpora in that two lemmata co-occur with the progressive in the later dataset but not in BNC1994DS, and six that occur in the earlier dataset but no longer in BNC2014S. Looking at frequency patterns of individual lexical verbs, HURT presents an interesting case as it occurs in all four possible combinations of dynamic and stative readings on the one hand, and the progressive and the simple on the other. Overall, the stative simple (e.g. *my tummy* ***hurts***, BNC2014S, S8X7) is the most commonly used combination, while the dynamic uses remain in the minority and show a decreasing trend between the datasets.

Using a distinctive collexeme analysis (DCA, see e.g. Gries & Stefanowitsch, 2004), we contrast dynamic uses of stative verbs with the simple/non-progressive use of these verbs. Lemmata that are statistically significantly attracted to the progressive include LOOK, HANG and HOLD; the collostructional strength has become weaker for LOOK and stronger for HANG and HOLD, indicating subtle changes in the use of dynamic stative progressives over time. With regard to the simple construction, we find that the stance verbs SIT and STAND are attracted to the simple (e.g. *he* ***sat down*** *next to Louise,* BNC1994DS, PS08X) and that the attraction has become stronger over time. Interestingly, Rautionaho and Fuchs (2020) show a diverging pattern for stance verbs occurring with ‘stative’ stative verbs - for them, the attraction is to the progressive (e.g. *so you****’re sitting*** *there*, BNC2014S, S7TT). Further research is necessary to investigate these patterns in detail.

Finally, our third research question concerns the possible effect dynamic uses of stative verbs may have had on the impression of increasing use of stative progressives, which has been suggested as one of the catalysts for the overall increase in the use of the progressive construction. Although we find an increasing trend in the present data, it is not significant enough to affect the results on stative uses of stative progressives reported in Rautionaho and Fuchs (2020); when the two datasets are combined, there is no increase over time (BNC1994DS 6.5% vs. BNC2014S 6.8%, LL=1.37, ns.). Thus, it is unlikely that the dynamic uses of stative verbs would have affected the use of (stative) progressives in recent BrE on a larger scale. Whether dynamic uses were indeed included or excluded in previous studies remains unclear, however, which leaves the question regarding change before the 1990s open. Overall, our two studies show that both stance verbs and dynamic uses of stative verbs should be separated from 'stative' stative verbs.

*References*

Gries, S. Th. (2014). Coll.analysis 3.5: A script for R to compute perform collostructional analyses.

Gries, S. Th., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on ‘alternations’. *International Journal of Corpus Linguistics, 9*(1), 97-129.

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009)*. Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319-344.

Mair, C. (2006). *Twentieth-century English: History, variation and standardization*. Cambridge: Cambridge University Press.

Rautionaho, P., & Fuchs, R. (2020). Recent change in stative progressives: A collostructional investigation of British English in 1994 and 2014. *English Language and Linguistics, 1*, 1-26.

Smitterberg, E. (2005). *The progressive in 19th-century English: A process of integration*. Amsterdam: Rodopi.

## Gentens, Caroline

***Stockholm University***

**caroline.gentens@english.su.se**

## Höglund, Mikko

***Stockholm University***

**mikko.hoglund@english.su.se**

Type of Contribution: Full Paper

# From manner to fact

Diachronic shifts towards propositional *the way that*-clauses

*Abstract*

In PDE, complex NPs with the noun *way* can introduce a factive (i.e. non-reported) propositional complement as in (1). In these contexts, the complex NPs with *way* do not necessarily invoke a manner reading which fills an adverbial gap in the following clause (cf. How *has he involved the rhythms of the branches__while keeping the forms in their place*?); rather, they are functionally similar to clauses introduced by *the fact that* (Legate, 2010).

(1) I liked Anthony Eyton's 'Oak Tree', and I specially admired **the way that** he has involved the rhythms of the near branches with the top of the oak in the middle-distance while keeping the forms in their place. (OED n. *rhythm*, 1965)

In other uses, the noun *way* is the antecedent of a relative clause, in which it functions as a path (2) or manner (3) expression.

(2) but euery day she wente and loked and espied **the waye that** he shold come yf she myght see hym come fro ferre (EEBO, 1483)

(3) he trusteth that she loueth hym the more Wherof he hath a pyteous herte and mynde and cannot be in rest tyll that he fynde **the ways how** he may her content &; ease all that he can (EEBO, 1509)

Historically, the first uses of *the way* + finite clause mainly involve path expressions in which the NP functions as an adverbial or complement (2), especially to predicates expressing some kind of motion (e.g. *tread*, *go, keep*). In Early Modern English the path expressions increasingly make way for a use in which *way* serves as an adverbial manner expression within the relative clause (3). Only in a later stage, in Late Modern English, we can see cases where the sense of manner modification to the relative clause can really be lost (1).

In this paper, we present a corpus-based study of the semantic changes affecting *the way that*-clauses. The analysis is based on extractions for (spelling variants of) the noun *way* from the Early Modern (EEBO, PPCEME2) and Late Modern (CEAL, PPCMBE1) English periods, with a focus on the pattern 'verb + the + way + clause'.

The data from Early and Late Modern English corpora show that in the Early Modern period the path usage of *the way that*-clauses was clearly the most frequent one. Often in religious texts the path was seen not as a literal, concrete path, but as a metaphorical path leading to heaven (or hell). These more metaphorical path uses seem to have functioned as a basis for the development of manner uses of *the way that*-clauses. In the Late Modern data, manner uses of *the way that*-clauses are attested more fequently than path-uses, and in addition, towards the end of the 19th century, the complementizer-like uses of *way* begin to appear in the data.

The historical development attested for *way* sheds light on grammaticalization on two different levels: more specifically on the grammaticalization patterns of *way* (see e.g. Tabor & Traugott, 1998 for *anyway*), and on a more general level on the mechanisms underlying a cross-linguistic tendency for path and manner expressions originating in lexical items to grammaticalize into clause linkers such as conjunctions introducing complement clauses (Boye & Kehayov, 2016; Güldemann, 2008; Hopper & Traugott, 2003 [1993]). Moreover, it provides a possible explanation for the semantic specialization of the complementizer-like use of *way*, i.e. for the fact that it is used to introduce factive, but not reported complements in PDE, cf. ?*He said the way that he balanced the rhythms of the branches in the painting.*

*References*

Boye, K. & P. Kehayov (eds). (2016). *Complementizer semantics in European languages*. Berlin: Mouton.

Güldemann, T. (2008). *Quotative indexes in African languages: A synchronic and diachronic survey*. Berlin: Mouton de Gruyter.

Hopper, P. J., & Traugott, E. C. (2003 [1993]). *Grammaticalization*. 2nd edition. Cambridge: Cambridge University Press.

Legate, J. A. (2010). On how *how* is used instead of *that*. *Natural Language and Linguistic Theory, 28*, 121-134.

Tabor, W., & Traugott., E. C. (1998). Structural scope expansion and grammaticalization. In R. A. Giacalone & P. J. Hopper (Eds.), *The limits of grammaticalization* (pp. 229-272). Amsterdam: John Benjamins.

*Data sources*

CEAL: Corpus of Early American Literature

EEBO: Early English Books Online.

OED: Oxford English Dictionary

PPCEME2: Penn-Helsinki Parsed Corpus of Early Modern English

PPCMBE1: Penn Parsed Corpus of Modern British English

**Gilquin, Gaëtanelle**

***UCLouvain***

**gaetanelle.gilquin@uclouvain.be**

**Granger, Sylviane**

***UCLouvain***

**sylviane.granger@uclouvain.be**

Type of Contribution: Full Paper

# The passive and the lexis-grammar interface

An inter-varietal perspective

*Abstract*

The passive has traditionally been seen as a purely grammatical phenomenon, resulting from the transformation of an active sentence. Yet, some voices have arisen to show (i) that the distinction between the active and the passive is not a strict dichotomy, but a gradient (see Svartvik, 1966; Quirk et al., 1985, p. 167ff.), and (ii) that the passive is a multifaceted construction, displaying morphological and syntactic attributes, but also discourse effects, as well as stylistic, lexical and phraseological preferences (e.g. Biber et al., 1999; Granger, 2013). This view, however, is still largely absent from the field of second language acquisition.

This paper aims to investigate the passive, seen as a gradient phenomenon at the lexis-grammar interface, among different EFL and ESL populations. It seeks to answer four research questions:

RQ1 Do learners of English use the passive in a native-like manner in terms of frequency?

RQ2 Do learners of English use the passive in a native-like manner in terms of lexical preferences?

RQ3 Do learners of English use the passive in a native-like manner in terms of phraseological sequences?

RQ4 Do ESL learners use the passive differently from EFL learners (in terms of frequency, lexical preferences and phraseological sequences)?

The data used in this study come from the latest version of the International Corpus of Learner English (ICLEv3; Granger et al., 2020) and represent four EFL populations (French-speaking Belgian students [FR], German students [GE], Korean students [KR] and Serbian students [SE]) and four ESL-like populations (Chinese-speaking Hong Kong students [HK], Dutch-speaking students from the Netherlands [DU], Norwegian students [NO] and Tswana students [TS]). These data are compared with native data taken from the Louvain Corpus of Native English Essays (LOCNESS). The methodology itself eschews a strictly grammatical definition of the passive by not only searching for the BE auxiliary followed by a past participle, but also including cases where BE is followed by a form tagged as an adjective in *-ed* (e.g. BE *concerned*, BE *involved*). In total, twenty forms were investigated: *concerned*, *considered*, *described*, *found*, *given*, *interested*, *involved*, *known*, *made*, *needed*, *provided*, *satisfied, said*, *seen*, *shown*, *suggested*, *supposed*, *taken*, *used* and *written*.

After the manual disambiguation of all the hits, we ended up with 4,733 occurrences of the passive. On the basis of these data, we computed, for each verb and each population, the relative frequency of the passive, as well as the passive ratio, which corresponds to the proportion of passive uses out of all the occurrences of the lemma and thus shows the degree to which a verb is attracted to the passive voice. A more qualitative analysis was also carried out that examined the phraseological patterns emerging from the data.

The frequency analysis reveals that the passive, overall, is underused by learners, but that the frequency varies quite widely among the L1 populations (from 167 occurrences per 100,000 words among the Korean learners to 310 among the Dutch learners). The passive ratio also shows some variation, both between the verbs and the populations. While the top three is the same in LOCNESS and ICLE, a number of verbs have a significantly lower passive ratio in ICLE, namely CONSIDER, MAKE, NEED, SEE, SHOW, USE, whereas only one has a significantly higher passive ratio, namely DESCRIBE. L1-specific results include the higher passive ratio of CONCERN in ICLE-FR and the lower passive ratio of INVOLVE in ICLE-HK. A hierarchical cluster analysis based on the passive ratios for each verb in LOCNESS and in the different ICLE subcorpora revealed that, contrary to expectations, EFL and ESL populations do not cluster in two separate groups and that the varieties that cluster with LOCNESS and hence resemble the native variety most closely are ICLE-DU and ICLE-NO (two ESL-like varieties) as well as ICLE-GE and ICLE-FR (two EFL varieties).

Some of the quantitative results can be explained by the underuse or overuse of specific phraseological patterns or the presence of unidiomatic phrases. The higher passive ratio of *CONCERN* in ICLE-FR, for example, is the consequence of the overuse of the pattern *as far as X BE concerned*, illustrated by:

(1) *As far as Europe is concerned*, what we can observe is a marked tendency to specialization (ICLE-FR)

Impersonal *it*-patterns with a passive voice are also interesting to examine. Some of them are underused in comparison with native English. This is the case of the *it*-pattern with SAY in ICLE-FR, which amounts to 27%, as against 46% in LOCNESS. This result should be interpreted in the light of the high frequency of the active counterpart, *we can/could say that*, a literal translation of the French discourse formula *on peut/pourrait dire que*. Some impersonal *it*-patterns, on the other hand, are overused in comparison with native English which, as exemplified in (2) and (3), is often due to the presence of unidiomatic phrases whose use can be described as midway between Sinclair's (1991) 'open-choice principle' and 'idiom principle':

(2) *It can be well known* that the target of study is to gain knowledge in the school (ICLE-HK)

(3) *It is wide known* that university degree is the first and basic step in the life of the future (ICLE-SE)

The analysis makes it possible to answer the research questions as follows:

RQ1 Learners of English have an overall tendency to underuse the passive in comparison to native writers.

RQ2 A majority of the verbs under study display similar passive ratios in native and non-native English, but some verbs are significantly more or significantly less attracted to the passive among certain learner populations.

RQ3 Next to some perfectly idiomatic phraseological sequences used in native-like proportions, we find in the L2 data phraseological teddy bears as well as certain unidiomatic phrases which seem to be the result of both the ‘open-choice principle’ and the ‘idiom principle’.

RQ4 Overall, EFL and ESL appear to display similar tendencies, with a few differences that however do not go in the expected direction of ESL learners using the passive in a more native-like manner than EFL learners.

By bringing to light the very important role that lexico-grammar plays in passive constructions, this study shows that more efforts should be channelled into the teaching of phraseological aspects of the passive, and the development of tools and methods that can facilitate the acquisition of such aspects.

## *References*

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Essex: Pearson.

Granger, S. (2013). The passive in learner English: Corpus insights and implications for pedagogical grammar. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, Vol. 1* (pp. 5-15). Kobe University, Japan.

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English, version 3*. Presses universitaires de Louvain.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Svartvik, J. (1966). *On voice in the English verb*. Hague: Mouton.

**González Cruz, M. Isabel**

***Universidad de Las Palmas de Gran Canaria***

**isabel.gonzalezcruz@iulpgc.es**

Type of Contribution: Poster Presentation

## Building a corpus of Anglicisms in digital newspaper headings

*Abstract*

The influence of English on other languages is an engaging field of research which has attracted the attention of scholars and laypeople alike for decades, generating an enormous body of works. These works have proved the increasing process of Anglicization in most European communities and beyond. Anglicisms pervade every area of our daily life (Luján-García, 2012) and have been attested in all European languages (Furiassi, Pulcini & Rodríguez-González, 2012). Computers (Pano, 2007; Bolaños & Luján, 2010), economy (López-Zurita, 2005), sports (Rodríguez-González, 2012), fashion and beauty (Balteiro, 2014), TV-advertising (García-Morales et al, 2016), or leisure (González-Cruz, 2015) are some of the many areas where Anglicisms abound, thence spreading into more general spheres. All these areas tend to be covered by most newspapers in their various sections. In fact, newspapers recognizedly reflect current linguistic usages, playing a key role in their diffusion in various national settings, particularly in Spain, (Morín, 2006; González-Cruz, 2012; Nuñez-Nogueroles, 2018). This justifies the use of the press as a suitable source to examine the growing presence of Anglicisms in their headings, even more so with the current impact of online journalism.

This poster summarizes the results of an investigation aimed at compiling a corpus of Anglicisms used in the headings of digital press. The first stage of this piece of research was related to a short project funded by the Canary Islands' Government (grant CEI2018-32). Here a team of ULPGC researchers analyzed the main digital papers published today in the Canaries (Spain) between March 1st and June 1st 2019. I focused my attention on the regional paper, *Canarias 7,* whose headings revealed the occurrence of a large

number of Anglicisms, with a high percentage of proper nouns. Between September 7th 2019 and April 7th 2020 I developed a second unfunded stage, to continue the compilation and study of headings with Anglicisms from the same digital paper. As for the methodology, in both stages I registered all headings (also sub-headings) with Anglicisms in an Excel file, including details such as date of occurrence, paper section, type of Anglicisms, typographical marks and any other suitable observation related to their usage, forms and functions.

Due to space and time limits, here I will just focus on the results of this second stage, which confirm the significant presence of English proper nouns in a corpus of 692 headings with three general types of Anglicisms, namely, proper nouns; new or non-registered Anglicisms (i.e., those that do not appear yet in the DRAE, the official dictionary published online by the *Royal Academy of the Spanish Language*); and registered Anglicisms. Specifically, I found 405 headings with at least one of the 202 different proper nouns collected. These were classified into the following categories: (a) titles of films, plays, songs, TV channels, programs, names of Apps, social networks and publications; (b) names of social musical events; (c) names of shops, ships, hotels and enterprises; (d) names of sports, sport teams, gyms, events and tournaments; (e) names of organizations, celebrations, campaigns, challenges and prizes; (f) names of characters, singers or musical groups; (g) toponyms and leisure places; (h) trade marks (i) names related to Politics, and (j) Acronyms. Some of the many proper nouns found were names of English-speaking celebrities, mainly actors and actresses, but these were neither collected nor considered for our frequency count.

Other than proper nouns, the total number of different Anglicisms found in this second stage amounts to 146. In particular, the quantitative analysis of this data shows the occurrence of 90 headings with a total of 67 new Anglicisms and 207 headings using 79 registered Anglicisms. All these Anglicisms can also be further classified, following Furiasi, Pulcini and Rodríguez-González's (2012) typology, into 'unadapted' (eg. *influencer, top model, online, golf, topless, drag queen);* 'adapted' (*fútbol, básquet, parquin, tique, estrés, jáquer, selfi);* hybrids (*Sitycleta, Black Fraude, Jandíabike)* and 'pseudo-Anglicisms' (*balconing, bunkering, Vueling*).

Finally, below I will list some interesting observations about the data obtained:

(a) As expected, most Anglicisms (new and registered) work as nouns, followed far by adjectives, all maintaining their original categories in English, with a few exceptions of derived forms which provoke cases of verbalization or nominalization. Eg.: from ***rap*** *(N) > rapear (Verb*); from ***hacker*** *(N) > hackear* (Spanishized into *'jaquear') (Verb);* from ***leader*** *(Spanishized into 'líder') > liderar (Verb*); from ***reset*** *(Verb) > reseteo (N).*

(b) There seems to be a preference for English spelling, even in registered Anglicisms that have officially been adapted to Spanish spelling rules, such as básquet /*basket;* clic/*click;* stand/*stand;* estrés/*stress;* jaquear/*hackear;* parquin/*parking;* roquero/*rockero;* tique/*ticket.*

(c) Some English plural forms are often avoided, thus provoking grammatical disagreement, as in *"*unos **test," "*unos **youtuber," "*algunas *big *band".*

(d) New word formations with e(-) abound, eg., eBiblio, e-Sports, e-cigarrillos.

(e) Some new Anglicisms are created by adding English or Spanish suffixes to already registered Anglicisms: *coach(ing); swing(ers); rap(ea); surf(ero); look(azos).*

(f) Some compound Anglicisms are only partially registered: *(baby) boom, (talent) show*.

(g) Some Anglicisms registered in *DRAE* are used but with other meanings: *baby, top.*

(h) Most Anglicims perform a ***referential function***, with a few instances of the ***expressive function***, such as the headings "BOOOOM!" or "Welcome, Angelina!"

The compilation will continue shortly with a third stage, which will allow us to build and study a larger and updated corpus of Anglicisms in use in today's Canarian-Spanish press.

## *References*

Balteiro, I. (2014). The influence of English on Spanish fashion terminology: -ing forms. *ESP Today, 2*(2), 156-173.

Bolaños-Medina, A., & Luján-García, C. (2010). Análisis de los anglicismos informáticos crudos del léxico disponible de los estudiantes universitarios de traducción. *Lexis, 34*(2), 241-274.

Furiassi, C., Pulcini, V., & Rodríguez-González, F. (Eds.). (2012). *The anglicization of European lexis.* Amsterdam, Philadelphia: John Benjamins.

García-Morales, G., María Isabel González Cruz, M. I., Luján García, C., & Jesús Rodríguez Medina, M. (2016). *La presencia del inglés en la publicidad televisiva española, 2013-2015.* Madrid: Síntesis.

González-Cruz, M. I. (2015). Anglizicing leisure: The multimodal presence of English in Spanish tv adverts. *Calidoscópio, 13*(3), 339-352

González-Cruz, M. I. (2012). English in the Canaries: Past and present. *English Today. The International Review of the English Language, 109*(1), 20-28.

González-Cruz, M. I., & Jesús Rodríguez-Medina, M. (2011). On the pragmatic function of Anglicisms in Spanish: A case study. *Revista Alicantina de Estudios Ingleses, 24*, 257-273.

López-Zurita, P. (2005). Economic Anglicisms: Adaptation to the Spanish linguistic system. *Ibérica, 10*, 91-114.

Luján-García, C. (2012). The impact of English on Spanish daily life and some pedagogical implications. *Nordic Journal of English Studies, 11*(1), 1-21.

Morín, R. (2006). Evidence in the Spanish language press of linguistic borrowings of computer and Internet-related terms. *Spanish in Context*, *3*(2), 161-179.

Nuñez-Nogueroles, E. E. (2018). A corpus-based study of Anglicims in the 21st century Spanish press. *Analecta Malacitana Electrónica, 44*, 1-37.

Pano, A. (2007). Los anglicismos en el lenguaje de la informática en español. El 'misterioso mundo del tecnicismo' a través de foros y glosarios en línea. Retrieved from: http://amsacta.unibo.it/2370/1/Lenguaje_informatica_ceslic_PANO.pdf

Rodríguez-González, F. (2012). Anglicismos en el mundo del deporte: variación lingüística y sociolingüística. *BRAE*, tomo XCII, cuaderno XXXVI, 261-285.

**Gorlach, Marina**

***Metropolitan State University of Denver***

***gorlach@msudenver.edu***

Type of Contribution: Full Paper

**To not let it happen or not to let it happen?**

Corpus-based analysis of negative infinitive alternation in discourse

*Abstract*

The cognitive complexity of negative meaning underlies the various lexical, syntactic, and pragmatic ways of expressing it in English. This paper discusses the structural expressions of the negative infinitive in various types of discourse. The focus is on the alternation between the two forms of the negative infinitive, *not to VERB* vs. *to not VERB*, as they are distributed across written and spoken texts of different genres. Only one of them, *not to VERB*, is recognized as existing in English by grammarians. The corpus data, however, demonstrate that the so-called split negative infinitive is actively and more frequently used in certain types of discourse.

This is a corpus-based study making the connection between the forms and meanings of the negative infinitive constructions themselves, as well as in relation to genre, register, and discourse situations. The paper presents the comparative frequency of using each form in discourse and discusses the pragmatic implications and applications of the observed distribution. As per COCA (Corpus of Contemporary American English) database, the frequency of *to not VERB* constructions is significantly higher in the spoken register (13.33 per million) as opposed to academic texts (4.10 per million). The newspaper genre usage is second high (7.02 per million), while the frequency in fiction and magazine genres is similar to academic texts (4.24 and 4.45 per million, respectively). Chronologically, the frequency of *to not VERB* shows a steady growth each decade since 1990, rising from 4.56 per million quotations in 1990-1994 to 9.59 per million in 2010-2015.

The paper explores the relationship between the form and communicative function, as well as the role of the additional lexical and non-lexical devices in generating the negative meaning. Theoretically and methodologically, this analysis relies on the concept of markedness and non-random distribution, treating split negative infinitive as a marked item in discourse. Since the two negative infinitive constructions co-exist, the speakers are presented with a subconscious choice, which is contextually motivated. The non-synonymy assumption claims that different signals/forms indicate different meanings/messages, however subtle such a difference may be, and that discourse situations and communicative goals serve as a motivating factor.

(1) KING: Bob Grant, what can you tell us? GRANT: Well, I would just caution folks *not to lose confidence*. (CNN, 1997)

(2) After 9/11, U.S. President Bush asked Americans to carry on with their lives, *to not lose confidence*, and to continue spending. (Murray, 2013)

The study takes a closer look at specific contexts and meanings of the negative infinitive considering the criteria that go beyond the formal linguistic aspects of language, such as the communicative situation and interplay between the interlocutors. As follows from the corpus-based analysis, one of the constructions is significantly less frequent overall, but has shown dominance in certain types of discourse and a significant increase in frequency over the last decade.

## *References*

Brinton, L. (2017). *The evolution of pragmatic markers in English: Pathways to change.* Cambridge: Cambridge University Press.

Conrad, S., & Biber, D. (2009). *Real grammar: A corpus-based approach to English*. New York: Pearson Longman.

Napoli, E. (2006). Negation. *Grazer Philosophische Studien, 72*(1), 233-252.

Tobin, Y. (1990). *Semiotics and linguistics*. London: Longman.

**Gut, Ulrike**

***University of Münster***

**gut@uni-muenster.de**

**Li, Zeyu**

***University of Münster***

**zeyu.li@uni-muenster.de**

Type of Contribution: Full Paper

## The /ʍ/-/w/ contrast in Scottish Standard English

A corpus-based approach to sound variation and change

*Abstract*

The voiceless labial-velar fricative /ʍ/ (also represented as /hw/) is a consonant that is generally described as being part of the Scottish English phoneme inventory (Wells, 1982, p. 408; Giegerich, 1992, p. 36; Jones, 2002, p. 27; Stuart-Smith, 2008, p. 63). It is used to pronounce the digraph <wh-> in many words such as *which*, *where* and *whether*, which are pronounced with an initial /w/ in British Standard English, resulting in Scottish English minimal pairs like *which* vs. *witch*, *where* vs. *wear*, and *whether* vs. *weather*.

/ʍ/ exists in Scottish English because, like in some varieties of Irish, American and Canadian English, the so-called *wine-whine* merger did not take place, in which historical Old and Middle English /hw/ was replaced by /w/. However, several studies have suggested that this historical contrast is weakening for an increasing number of Scottish English speakers in the urban areas of Glasgow (Stuart-Smith, 2008, p. 63), Edinburgh (Schützler, 2010) and Aberdeen (Brato, 2007, 2014). Significant differences between older and younger Scottish speakers have been found for the towns of Livingston (Robinson, 2005), Glasgow (Timmins et al., 2004), Aberdeen (Brato, 2014) and Edinburgh (Schützler, 2010): the younger speakers generally are not sensitive

to the contrast of /ʍ/ and /w/ and use both interchangeably for <wh->, while older speakers produce fewer /w/ in this context.

The realization of the /ʍ/-/w/ contrast further appears to be influenced by the social factors class and gender: In Glasgow, middle-class speakers, both adolescents and older speakers, use /ʍ/ more frequently than working-class speakers. Especially middle-class men strongly favour the /ʍ/ variant, while working-class boys and girls show a very low usage of /ʍ/ (Lawson and Stuart-Smith, 1999; Stuart-Smith et al., 2007; Timmins et al., 2004). This distribution with class was also found by Brato (2014) for Aberdeen speakers. Schützler (2010) furthermore observed gender differences with a greater erosion of the /ʍ/ - /w/ contrast for male than for female speakers in Edinburgh.

In addition to sociolinguistic predictors, Schützler (2010) and Brato (2014) also found language-internal factors affecting the realization of the /ʍ/ - /w/ contrast. Their findings show that the /ʍ/ variant favours stressed syllables, word-initial and postpausal position, which is explained by “the slightly more effortful /ʍ/” being more easily pronounced, least influenced by coarticulatory factors and receiving more attention with a preceding pause rather than in unstressed or utterance-medial position (Schützler, 2010, p. 15; Brato, 2014, p. 40).

However, little is known about whether the /ʍ/-/w/ contrast is still maintained in supraregional Scottish Standard English (SSE) and whether the influencing factors found for various types of urban Scottish English also determine its presence and distribution in the Scottish standard variety. While many authors claim that the /ʍ/-/w/ contrast is present in contemporary SSE (Giegerich, 1992, p. 36; Jones, 2002, p. 27; Stuart-Smith, 2008, p. 63), to the best of our knowledge, no empirical studies have been carried out so far to substantiate this. It is thus the aim of this study to explore, based on a large phonological corpus of SSE, whether the /ʍ/-/w/ contrast still exists and whether its presence and distribution are influenced by the social factors age and gender as well as the language-internal factors stress and phonetic contexts.

Some first evidence for a larger presence of /ʍ/ in the standard variety of Scottish English might be deduced from those studies of urban Scottish English that compared two or more speaking styles differing in the level of formality. Some of them found that in the most formal speaking style, i.e. word lists, in which the speakers' target is presumably closest to the standard form due to a high level of attention given to the pronunciation of each word, more /ʍ/ are produced for <wh-> than in less formal styles such as passage reading and conversations (Brato, 2007; Stuart-Smith et al., 2007). Stuart-Smith et al. (2007) further found that the social factors class, age as well as age*gender only significantly influenced the realization of the /ʍ/-/w/ contrast in Glaswegian speech in the conversations but not the word lists. In our study, we will therefore test the following predictions:

- /ʍ/ is largely present for <wh-> in Scottish Standard English
- the social factors age and gender will play a smaller role in determining the presence and distribution of [ʍ] than the phonetic factors and stress

The spoken data were drawn from ICE Scotland, the Scottish component of the International Corpus of English (ICE) (Schützler et al., 2017) and comprised several categories of formal speech and dialogues[1], produced by 182 middle-class speakers (75 females, 107 males, aged between 17 and 78) from all over Scotland. A total of 1,177 <wh> tokens were analyzed auditorily by two independent coders. Moreover, the degree of acoustic periodicity of each token was measured by extracting the median of harmonicity (also referred to as harmonics-to-noise ratio, see e.g. Boersma, 1993, Hamann & Sennema, 2013) using Praat (Boersma & Weenink, 2017). For the statistical analysis, linear mixed-effects regression models with MEDIAN OF HARMONICITY as dependent variable, and SPEAKER and WORD as random intercepts were used. Fixed predictors including the social factors AGE and GENDER and the internal factors preceding/following CONTEXT and STRESS were tested. The models were fitted using the R-package {lme4} (Bates et al., 2015).

The results of the auditory analysis show that /ʍ/ and /w/ are not fully merged for about half of the speakers of SSE: 36% of all <wh> tokens were realized as [ʍ]. 41% of all analyzed speakers produced exclusively [w] and thus

[1] Text categories include broadcast discussion, broadcast talk, broadcast interview, broadcast news, cross examination, legal presentation, non-broadcast talk, parliamentary speech, unscripted speech.

seem to have a complete /ʍ/-/w/ merger, while about 10% of the speakers produced [ʍ] for <wh> categorically in their speech. The remaining 49% of the speakers produced both [w] and [ʍ] for <wh>. Concerning the acoustic properties of the /ʍ/-/w/ contrast, the <wh> tokens classified as [w] (*M* = 8.39, *SE* = 0.21) in the auditory analysis have a considerably higher median of harmonicity than those labelled as [ʍ] (*M* = 4.89, *SE* = 0.22), *t*(1175) = -10.9, *p* < .001.

The mixed-effects linear regression models on the median of harmonicity showed that the realization patterns of the /ʍ/ variant is conditioned by both social and language-internal factors. In terms of social effects, both age and gender showed statistically significant impact on the realization of [ʍ], but not their interaction age*gender. More specifically, older speakers and males produced a lower median of harmonicity (and thus more [ʍ]) than younger speakers and females. Regarding internal factors, preceding phonetic context plays a statistically significant role with the realization of [ʍ] favouring word-initial postpausal position the most and word-internal position the least. Phonological stress and following vowel environment, however, did not reach statistical significance.

To summarize, the present study found that /ʍ/ for <wh> is still present in SSE, albeit with lower rate than claimed in Giegerich (1992, p. 36), Jones (2002, p. 27) and Stuart-Smith (2008, p. 63). The variable merger patterns across individual speakers were shown to be conditioned by both social (age and gender) and language-internal factors (preceding phonetic context), with the /ʍ/ variant being favoured by older speakers, males, and at the beginning of an utterance. Contrary to our assumption, language-internal factors do not play a greater role than social factors in influencing the rate of [ʍ] production by SSE speakers.

### *References*

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., & Bolker, M. B. (2015). Package ‘lme4’. *Convergence*, *12*(1), 2.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences (University of Amsterdam),* 17, 97-110.

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer program]. Version 6.0.31. Retrieved from: http://www.praat.org/

Brato, T. (2014). Accent variation and change in north-east Scotland: The case of (hw) in Aberdeen. In R. Lawson (Ed.), *Sociolinguistics in Scotland*, Houndmill (pp. 32-51). New York: Palgrave Macmillan.

Giegerich, H. (1992). *English phonology: An introduction*. Cambridge: Cambridge University Press.

Hamann, S., & Sennema, A. (2013). *Acoustic differences between German and Dutch labiodentals*. Universitätsbibliothek Johann Christian Senckenberg.

Johnston, P. (1997). Regional variation. In C. Jones (Ed.), *The Edinburgh history of the Scots language* (pp. 433-513). Edinburgh: Edinburgh University Press.

Jones, C. (2002). *The English language in Scotland: An introduction to Scots*. East Linton: Tuckwell Press.

Robinson, C. (2005). Changes in the dialect of Livingston. *Language and Literature, 14*(2), 181-193.

Schützler, O. (2010). Variable Scottish English consonants: The cases of /ʍ/ and non-prevocalic /r/. *Research in Language,* 8, 5-21.

Schützler, O., Gut, U., & Fuchs, R. (2017). New perspectives on Scottish Standard English: Introducing the Scottish component of the International Corpus of English. In S: Hancil & J. Beal (Eds.), *Perspectives on Northern Englishes* (pp. 273-301). Berlin: Mouton de Gruyter.

Stuart-Smith, J. (2008). Scottish English: Phonology. In B. Kortmann & C. Upton (Eds.), *Varieties of English: The British Isles* (pp. 48-70). Berlin, New York: Mouton de Gruyter.

Timmins, C., Tweedie, F., & Stuart-Smith, J. (2004). Accent change in Glaswegian (1997 corpus): Results for consonant variables. Department of English Language, University of Glasgow.

Wells, J. (1982). *Accents of English*. Cambridge: Cambridge University Press.

Herzky, Jenny

*University of Bamberg*

jenny.herzky@uni-bamberg.de

Schützler, Ole

*University of Bamberg*

ole.schuetzler@uni-bamberg.de

Type of Contribution: Full Paper

# Contractions in negative clauses

## Comparing British-Isles Englishes

*Abstract*

Negative clauses constructed from a primary or auxiliary verb (e.g. BE, HAVE, WILL) and the negator *not* can be realized as (cf. Castillo González, 2007): (1) negative contractions (or *not*-contractions; **NC**s), (2) operator contractions (**OC**s) and (3) uncontracted negatives (**UC**s).

(1) [… W]e often see people saying Scotland isn't digitally forward looking. (ICE-SCO rep-40)

(2) […] Scotland's not digitally forward looking.

(3) […] Scotland is not digitally forward looking.

OCs have been found to be more likely (i) with pronominal subjects (*He's not here* vs. *My colleague isn't / is not here*), (ii) if the verb is an auxiliary, and (iii) at lower levels of formality (cf. Tagliamonte and Smith, 2002). Furthermore, Quirk et al. (1985; cf. Miller, 2008) consider NCs to be more common than OCs, but state that the latter may be more frequent in Northern and Scottish Englishes (cf. Miller, 2008). This is the point of departure for the present research.

We compare the frequencies of negative constructions of all three types in Scottish Standard English (SSE), Southern British Standard English (SBSE)

and Irish Standard English (IrSE). We expect that in all three varieties, language-internal and stylistic constraints operate as described above, but that OCs are generally more likely to occur in SSE.

Based on the Scottish, Irish and British (effectively: ‘southern British’) components of the International Corpus of English (ICE; cf. Kirk & Nelson, 2018), we extracted all instances of negated *is*, *are*, *have* and *will*. In IrSE and SBSE, all genres except conversations and telephone calls are represented. For the SSE data, the selection of genres is further reduced, due to the incomplete state of the corpus. To a large extent, our model design takes care of this issue, but we will also expand our database in future research. We made a total of *n* = 2,829 observations in *n* = 836 texts.

Data were coded for the outcome **TYPE** (NC, OC, UC), as well as the predictors **VERB** (*is*, *are*, *have, will*), **MODE** (spoken/written), **VARIETY** (SSE/SBSE/IrSE), and **SUBJECT** (pronoun/NP). The type of verb (auxiliary vs. full verb) was disregarded, since this predictor seems to lack theoretical motivation. Furthermore, the verb WILL displays categorical behaviour, i.e. it does not function as a full verb in present-day English.

A Bayesian multinomial mixed-effects regression model with random effects **TEXT** and **GENRE** was fitted to predict the choice of construction, using the R-package brms (Bürkner, 2020). It was specified thus: type ~ (verb + mode) * variety + subject + (verb|text) + (verb|genre).

Figure 1 shows that differences between the three varieties are not pronounced. OCs with *have* and *will* are virtually non-existent. Compared to the other two varieties, contractions in SSE are somewhat more likely with *is* (OCs and NCs) and *will* (NCs) and less likely with *have* (NCs). There is no evidence of a special role of OCs, however.

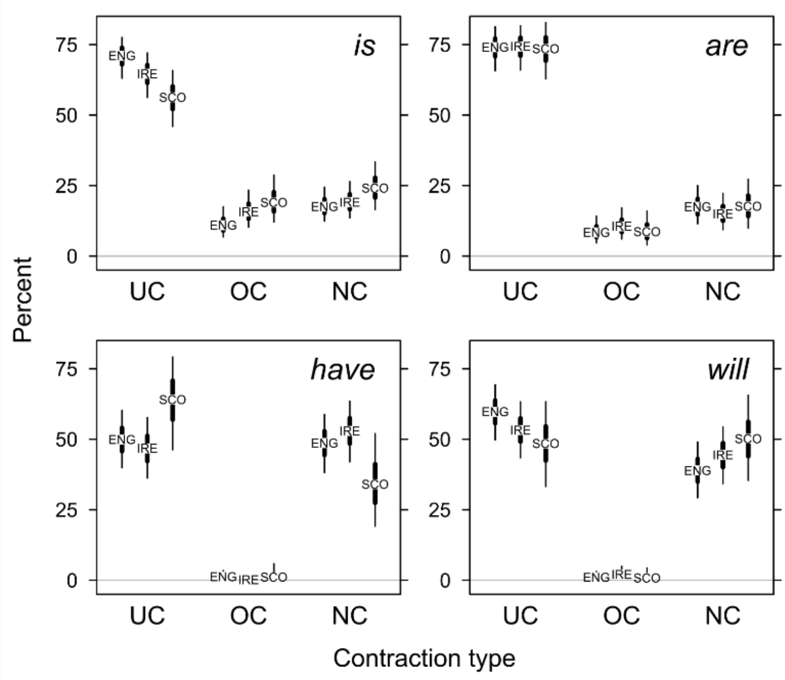


Figure 1: Percentages of construction types by variety and verb (50% and 90% uncertainty intervals).

Figure 2 focuses on the absolute percentage point differences between speech and writing. The general patterns are not only unsurprising, but they are also similar across varieties: UCs are much less likely in speech, and, conversely, both OCs and NCs occur at higher rates. It is striking that, with all four verb forms, SSE is least affected by this dimension of variation. However, this special behaviour (i.e. the lack of responsiveness to mode of production) surfaces only in connection with NCs.

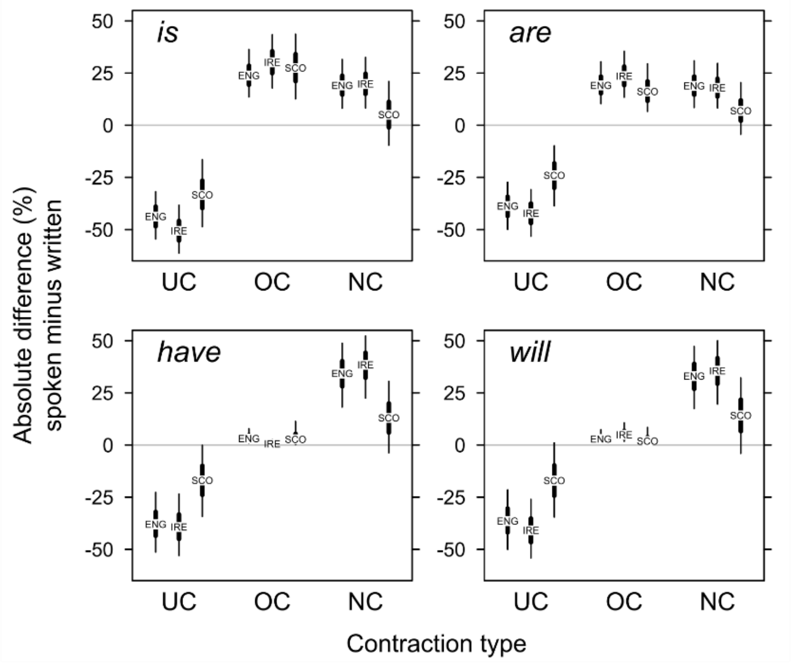


Figure 2: Differences between speech and writing (50% and 90% uncertainty intervals).

Beyond the effects discussed thus far, we find that pronominal subjects correlate with much higher rates of OCs and also have a weak favouring effect on the rate of NCs.

Our research thus far does not confirm that SSE is characterized by higher rates of OCs; the three British-Isles varieties are very similar in their use of the constructions we inspected. In our ongoing work on this topic, we will include more data from ICE-Scotland. Furthermore, we will add two language-internal factors to our analysis: (i) the length of the syntactic subject and (ii) the phonological connection between subject and verb. Concerning the former, we expect subjects that constitute longer and more complex NPs to be less likely to be of the OC type - some of this is already captured by effect of pronominal subjects, which are short by their very nature. Concerning the second point, we expect OC to be very unlikely if it results in a clash of homorganic sounds (e.g. *This's* good; *Bill'll do it*), even to the extent that such cases are nearly invariable and might need to be excluded. Our study is probably skewed to some extent by not considering these factors and will therefore be updated in due course.

## *References*

Bürkner, P. (2020). *brms: Bayesian Regression Models using ‘Stan’*. R-package version 2.12.0. Retrieved from: https://cran.r-project.org/web/packages/brms/brms.pdf

Castillo González, M. (2007). *Uncontracted negatives and negative contractions in contemporary English: A Corpus-based study*. University of Santiago de Compostela. PhD thesis.

Kirk, J., & Nelson, G. (2018). The International Corpus of English project: A progress report. *World Englishes, 34*(4), 697-716.

Miller, Jim E. (2008). Scottish English: Morphology and syntax. In B. Kortmann & C. Upton (Eds.), *Varieties of English: The British Isles* (pp. 299-327). Berlin: Mouton de Gruyter.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language.* London: Arnold.

Tagliamonte, S., & Smith, J. (2002). ‘Either it isn’t or it’s not’. neg/aux contraction in British dialects. *English World-Wide, 23*(2), 251-281.

**Hiltunen, Turo**

***University of Helsinki***

**turo.hiltunen@helsinki.fi**

**Vartiainen, Turo**

***Tampere University***

**turo.vartiainen@helsinki.fi**

Type of Contribution: Full Paper

## A corpus-pragmatic analysis of linguistic democratization in the British Hansard: Comparing the two Houses

*Abstract*

Linguistic democratization, understood as a group of discourse-pragmatic processes related to societal and sociocultural changes (Leech et al., 2009; Farrelly & Seoane, 2012), has recently been explored in parliamentary discourse in several studies using the *Hansard Corpus* (e.g. Spirling, 2016; Hiltunen et al., 2020). While democratization, colloquialization and related phenomena can be understood and operationalized in different ways, these studies have found evidence of a general progression towards increasingly democratic, colloquial or informal usage, although the precise patterns vary depending on a number of contextual factors, such as changes in the reporting conventions (Alexander & Dallachy, 2019; Hiltunen et al., 2020). The aim of the present paper is to provide a more detailed examination of the precise impact of these contextual factors on frequency data, and the extent to which this kind of data can be linked to specific discourse-pragmatic processes in the *Hansard Corpus* (Alexander & Davies, 2015). The main focus of our study is on the comparison between the House of Commons and the House of Lords, which display diverging patterns with respect to a number of high-frequency linguistic phenomena.

In order to explore the potential colloquialization of parliamentary debates, we investigate three linguistic features that have been connected to informal style and a high degree of speaker involvement in previous research (e.g. Biber, 1988): private verbs with first-person pronouns (e.g*. I think)*, progressive constructions, and zero *that*-clauses. Our data show that the pace of democratiuation/colloquialization differs significantly between the two houses: in the House of Lords, the frequency of private verbs and progressives rises steadily towards the 20th century, which supports the hypothesis that parliamentary discourse has become colloquialized over time. However, a closer look at the Commons data reveals that, after a substantial increase earlier in the 20th century, the frequency of these features has in fact decreased in recent decades. This suggests that there are other factors in addition to colloquialization that need to be considered in the analysis of the Hansard data. This conclusion is further supported by our finding that the proportion of *that*-deletion has in general decreased in both Houses over time - a result that runs contrary to the colloquialization hypothesis.

Generally, we find that although the pace and direction of change is different for private verbs and progressives in the two Houses, the data show considerable levelling in the most recent decades. This levelling may be an indication of a change in the editing process (cf. Alexander & Dallachy, 2019). Although the exact cause for the increased similarity of the texts remains uncertain, our results emphasize the need to treat the data from the two Houses separately, particularly in diachronic studies.

In addition to investigating the Hansard data from the perspective of colloquialization, we also examine a pattern where one of the most frequent private verbs, *think*, is used to convey a specific pragmatic meaning. Our data show that when a speaker in the House of Lords or the House of Commons uses *think* with a third-person subject pronoun and a coreferential subject in the complement *that*-clause (e.g. *he$_1$ thinks that he$_1$…*), they often disagree with the proposition in the complement clause. This disagreement may be overtly expressed, as in (1), or implied, as in (2).

(1) The Foreign Secretary gives the impression to everyone that **he thinks that he** has unlimited time. **He has not**. (Commons, 1979)

(2) **He thinks that he** and his friends are to have a monopoly of free speech, that he may talk as loudly and as provocatively as he pleases […] (Lords, 1918)

Interestingly, this result only obtains when the referent of *he* is another MP or another Lord: the pattern is used to imply disagreement in 69% of all cases in the Commons, and 61% in the Lords, when the referent is another politician. However, the proportion drops to 13% in the Commons and 20% in the Lords when the referent is an ordinary citizen. Although more research is needed to study the scope of the phenomenon, these results provide tentative support to the idea that the pattern is particularly typical of debates, where expressions of agreement/disagreement are highly relevant to the emerging discourse. The results of our study can be summarized as follows.

i. The rate and direction of language change may differ substantially between the House of Commons and the House of Lords. This suggests that the texts have been subjected to different editorial norms and conventions, which in turn merit close attention in analyses (cf. Alexander & Dallachy, 2019). The levelling seen in the most recent decades may be a result of a more centralized effort to edit the reports.

ii. Future studies using the Hansard Corpus should take the potential effect of the House into consideration in the interpretation of the results. Ideally, the data from the House of Commons should be studied separately from the House of Lords.

iii. From a qualitative perspective, our study has identified a pattern that is associated with the pragmatics of (dis)agreement: *he thinks that he*. While the scope and variants of this pattern remain to be fully investigated, our initial results suggest that this usage may be particular to debates: disagreement is only rarely implied when the referent of *he* is not a participant in the parliamentary debates.

## *References*

Alexander, M., & Dallachy, F. (2019). Historic Hansard 1805-2005: Two centuries of speech representation. ICAME40, Neuchâtel, Switzerland.

Alexander, M., & Davies, M. (2015-). Hansard Corpus 1803-2005. Retrieved from: http://www.hansard-corpus.org

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Farrelly, M., & Seoane, E. (2012). Democratization. In T. Nevalainen & E.C. Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 392-401). Oxford: Oxford University Press.

Hiltunen T., Räikkönen, J., & Tyrkkö, J. (2020). Investigating colloquialization in the British parliamentary record in the late 19th and early 20th century. In T. Hiltunen & L. Loureiro-Porto (Eds.), *New perspectives on democratization* (Special issue). *Language Sciences,* 79.

Leech, G., Hundt, M., Christian, M., & Smith, N. (2009). *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

Spirling, A. (2016). Democratization and linguistic complexity: The effect of franchise extension on parliamentary discourse, 1832-1915. *Journal of Politics, 78*(1), 120-136.

**Huang, Ding**

***Ruprecht-Karls-Universität Heidelberg***

**elainehd.dh@gmail.com**

Type of Contribution: Work-in-Progress Report

## Formulaic sequences in Early Modern English

A corpus-assisted historical pragmatic study

*Abstract*

This on-going PhD project compares the use of formulaic sequences (FSs) in Early Modern English (EModE) dialogues and letters. The study aims at defining FSs in a way that can be applied in research from a perspective of historical pragmatics. I do not use the definition in Wray (2002, p. 9) in my PhD project, because FSs being “stored and retrieved whole from memory” cannot be tested on native EModE speakers, nor via corpus data (Schmitt, Grandage & Adolphs, 2004). The study also aims at enhancing the methodology to identify FSs in historical texts. Although the corpus-based approach is gaining popularity in investigating FSs in historical texts, resulting in more genres examined, more types of FSs identified, and more representative results (e.g. Culpeper & Kytö, 2010), problems and challenges remain (Kohnen, 2002).

The research questions are as follows: What is the form of formulaic sequences in EModE communicational texts? What functions do they serve? How do they characterize different text-types?

Based on the characteristics of FSs described in previous studies (See summaries in Wray and Perkins, 2000; Wood, 2015), this study defines FSs as various types of multi-word units which have relatively fixed syntactic structure, form a semantic unit and serve as a conventional pairing of form, meaning and function.

This project follows the Construction Grammar theory and sees FSs as constructions, backed by the empirical evidence from Buerki (2016). Nevertheless, several critical factors make FSs stand out from other

constructions for investigation, namely the degree of abstraction, fixedness, and productivity; (I pay special attention to the fixedness of FSs and argue that it has different denotations in PDE and EModE due to their syntactic differences). Constructions are “conventional learned form-function pairings at varying levels of complexity and abstraction” (Goldberg, 2013, p. 17), ranging from, for example, the highly abstract and productive passive constructions to fully lexical, surface-level word sequences that have restricted productivity. FSs are those constructions positioned from the middle of the range to the more lexical and less productive end, i.e. idioms (filled, partially filled, and minimally filled) in Goldberg’s words.

This PhD project examines two corpora with a corpus-assisted, semi-automatic approach. The corpus of dialogues contains texts taken from the Corpus of English Dialogues (CED, 1560-1760), and the corpus of letters contains texts from the Parsed Corpus of Early English Correspondence (PCEEC, 1480-1681). The study abstracts only texts dated from the year 1560 to 1680 from the original corpora so that the two corpora are comparable. Texts are normalized with VARD 2.

To identification of FSs, I firstly retrieve lexical bundles (LBs) using the WordSmith tool. This study treats LBs as a measure to identify FSs instead of a type of formulaic language. LBs are repetitive word sequences that consist of at least three words and are functional (Biber et al., 2004). FSs have the same features, but more than that. This study also acknowledges the limitation of identifying FSs via LBs; because unlike LBs, FSs are mostly syntactic and semantic complete. That is why, in the next step, LBs shall meet specific criteria in order to qualify as FSs. Generally speaking, these criteria check if an LB has a form that can only be altered to a restricted degree, if it is syntactically and semantically complete, if its meaning is compositional, and if it represents a fixed form-function mapping.

### Data analysis

FSs will be grouped according to their functions, using the functional classifications by Biber et al. (2004): stance expressions, discourse organizers, reference expressions, and expressions that serve special conversational functions. Although these classifications are used initially to classify LBs in conversation and academic prose in PDE, this study believes that they can be used to classify FSs in EModE texts because Markus  (2018) has used these classifications to classify LBs in Bess letters written in EModE and FSs in this study are identified via LBs.

The project has so far made progress in four aspects. Firstly, I have spent many efforts on defining FSs and distinguishing FSs from other concepts about multi-word sequences, such as constructions, patterns, lexical bundles, collocations, etc. Secondly, I have prepared the corpora by selecting and normalizing texts. Thirdly, I have designed the procedure the criteria of retrieving FSs from corpora. Lastly, I have run several pilot studies using a corpus of Shakespeare’s plays to testify the feasibility of the method. The next step, which I am currently taking, is to retrieve FSs from the corpora of EModE communicational texts and classify them according to their functions.

Though there are yet not enough significant findings to answer the research questions, the process of normalizing corpora and pilot studies reveal several exciting phenomena. In private letters, abbreviated word sequences might be a helpful hint of FSs. Both the writer and the recipient must know what the abbreviations stand for. If certain shortened word sequences reoccur only in letters by specific authors, this might be subject to the habit of these authors, rather than a common practice in EModE. For example, in (1) *Y. l. f* stands for *your loving father*, and *T. B.* is the initials of the author’s name *Thomas Browne*. However, if the abbreviated word sequences are common in letters of various kinds in general, then they could be considered as FSs. For example, in (2) ***Lieut.-Col.*** is short for *Lieutenant Colonel*, a commonly known military rank.

(1) *Love & blessing to my daughter Browne & you all.* ***Y. l. f. T. B****.* (PCEEC, BROWNE,120.018.369-370)

(2) … *partly by the false informations of* ***Lieut.-Col.*** *Briddeman and …* (PCEEC, WHARTON,18.007.348)

Secondly, FSs might be the middle stage of compounding in word-formation. For example, *to morrow* (tomorrow), *with in* (within), *a sseure* (assure), *in treat* (entreat), *a fordeth* (affords), *a bout* (about), *after noone* (afternoon), *a sundre* (asunder), *ffor sakinge* (forsaking), *how beyt* (howebeit), *a fresh* (afresh), etc. However, since they are mostly two-word sequences, it could also be an orthographic change. This issue deserves further research as an independent project. In this project, I joined and normalized them into their modern spelling.

## *References*

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405.

Buerki, A. (2016). Formulaic sequences: A drop in the ocean of constructions or something more significant? *European Journal of English Studies, 20*(1), 15-34.

Culpeper, J., & Kytö, M. (2010). Early modern English dialogues: Spoken interaction as writing. Cambridge: Cambridge University Press.

Kohnen, T. (2002). Methodological problems in corpus-based historical pragmatics: The case of english directives. In K. Aijmer & Bengt Altenberg (Eds.), *Language and Computers, Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)* (pp. 237-247). Amsterdam: Rodopi.

Markus, I. (2018). *The linguistics of spoken communication in Early Modern English writing: Exploring Bess of Hardwick's manuscript letters*. Cham: Palgrave Macmillan.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 127-152). Amsterdam: John Benjamins.

Shakespeare Corpus. Compiled by Mike Scott. Retrieved from: https://lexically.net/wordsmith/support/shakespeare.html.

Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. London/New York: Bloomsbury.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language and Communication, 20*(1), 1-28.

**Iyeiri, Yoko**

***Kyoto University***

**yiyeiri@bun.kyoto-u.ac.jp**

**Fukunaga, Mariko**

***Kyoto University***

**fukunaga.mariko.45u@kyoto-u.ac.jp**

Type of Contribution: Full Paper

# A corpus-based analysis of negation in some 19th-century American missionary documents in Honolulu

*Abstract*

The present study forms part of a larger project based on some major documents left by members of the American Board of Commissioners for Foreign Missions (ABCFM). With the aim of exploring the language of 19th-century American missionaries, our research team has compiled the ABCFM Hawaii Corpus (hereafter Hawaii Corpus) by assembling selected documents from the collections of the Hawaiian Mission Children's Society Library. The corpus, which currently comprises approximately 653,100 words, includes journals, letters, and an autobiography written by eight members of the ABCFM:

Levi Chamberlain (1792-1849), 228,500 words (journals)
Lorrin Andrews (1795-1868), 24,100 words (journals)
Peter Johnson Gulick (1796-1877), 55,800 words (autobiography)
Dwight Baldwin (1798-1886), 139,900 words (journals)
Elisha Loomis (1799-1836), 29,300 words (journals)
Maria (Patton) Chamberlain (1803-1880), 69,500 words (journals)
Richard Armstrong (1805-1860), 24,500 words (journals)
Clarissa Chapman Armstrong (1805-1891), 81,500 words (journals and letters)

On the basis of this dataset, we have explored various aspects of negation. Our approach is corpus-based and quantitative.

One of the major findings of this study is that numerous examples of *do*-less negation, as in *they knew not*, are still observed fairly extensively in the Hawaii Corpus. While some examples occur within quotations from the Bible, a large number of them are attested in the written discourse of the authors. This supports some recent studies that note the existence of *do*-less negation in the 19th century (e.g., Curry, 1992; Iyeiri, 2004) and even in the 20th century (e.g., Yadomi, 2015). In the Hawaii Corpus, around 20% of the relevant examples illustrate *do*-less negation. Excluding from these statistics the three conservative verbs *have*, *know*, and *doubt*, we still obtain a proportion around 10% for *do*-less negation.

As a matter of fact, lexical *have* in the Hawaii Corpus occurs most frequently in *do*-less negation. This is interesting, since lexical *have* usually appears in *do* negation in 20th-century American English. On the other hand, Hirota (2020, pp. 130-134) shows that *do*-less negation was still common with lexical *have* throughout the 19th century in America, refuting the statement by Varela Pérez (2007, p. 225) that the shift to *do* negation was more or less complete with *have* in 19th-century American English. The present study supports Hirota (2020) by showing that the *have not* type (i.e. *do*-less negation) is the norm in the Hawaii Corpus with only a few exceptions.

Aside from this general trend, it is also noticeable that different authors present different linguistic features. Although the eight authors are all relatively well-educated and aware of their shared missionary aims in the same community, their use of negative constructions differs to some extent. This may be due to the use of different styles in different documents. We would like to single out Clarissa C. Armstrong in this relation, since her use of negation often deviates from the overall tendency in the Hawaii Corpus. She is, for example, one of the two authors who employ *do* negation at a larger rate than 95% when *have*, *know*, and *doubt* are excluded, while the corresponding rate in the entire corpus is around 90% as mentioned above. This may be indicative of the relatively colloquial nature of Clarissa’s writings, which include a number of documents in the letter form.

Various other aspects of negation in Clarissa's English support this inference. Clarissa uses negative clauses more frequently than many other authors in the Hawaii Corpus, and this also indicates the colloquial nature of her English. See Tottie (1981, p. 271), who points to the more frequent occurrence of negation in spoken than in written English. Also, Clarissa is one of the authors whose English tends to employ the negative adverb *not* relatively frequently instead of other negative items such as *no* and *never*. The frequent use of *not* is considered to be a feature of colloquial English (cf. Tottie 1988, p. 262; Iyeiri, et al., 2015).

Furthermore, Clarissa is among the small number of authors whose English shows the *neither* … *or* construction instead of *neither* … *nor*. The construction with *or* instead of *nor* was often ruled out in grammars in the late Modern English period (cf. Nevalainen, 2014, p. 33). Since relevant examples of this construction are not numerous in the whole dataset, their attestation in her writing may not constitute strong evidence, but is worth considering in combination with the other aspects of negation mentioned above.

As hitherto discussed, the Hawaii Corpus as a whole demonstrates the state of 19th-century American English, while at the same time it provides material suitable for historical sociolinguistic analyses, showing the variability of English among different authors in the same period.

## *References*

Curry, M. J. (1992). The *do* variant field in questions and negatives: Jane Austen's complete letters and *Mansfield Park*. In M. Rissanen, O. Ihalainen, T. Nevalainen, & I. Taavitsainen (Eds.), *History of Englishes: New methods and interpretations in historical linguistics* (pp. 705-719). Berlin, New York: Mouton de Gruyter.

Forbes, D. W., Kam, R. T., & Woods, T. A. (2018). *A biographical encyclopedia of American protestant missionaries in Hawai'i and their Hawaiian and Tahitian colleagues, 1820-1900*. Hawaiian Mission Children's Society.

Hawaiian Mission Houses (n.d.). *Hawaiian Mission Children's Society Library*. Retrieved from: http://hmha.missionhouses.org/collections/show/3

Hirota, T. (2020). Diffusion of *do*: The acquisition of *do* negation by *have (to).* In M. Kytö & E. Smitterberg (Eds.), *Late Modern English: Novel encounters* (pp. 117-142). Amsterdam: John Benjamins.

Iyeiri, Y. (2004). The use of the auxiliary *do* in negation in *Tom Jones* and some other literary works of the contemporary period. In I. Moskowich-Spiegel Fandiño & B. Crespo García (Eds.), *New trends in English historical linguistics: An Atlantic view* (pp. 223-240). Coruña: Universidade da Coruña.

Iyeiri, Y., Yaguchi, M., & Baba, Y. (2015). Negation and speech style in professional American English. *Memoirs of the Faculty of Letters, Kyoto University, 54*, 181-204.

Nevalainen, T. (2014). Variation in negative correlative conjunctions in 18th-century English. In K. Nakagawa, F. Shigenobu, O. Imahayashi, K. Ishikawa, Y. Ishizaki, T. Kawabata, K. Koguchi, Y. Makita, F. Matsubara, S. Ohta, H. Osaki, M. Sawada, M. Uchida, K. Wakimoto, T. Yanagi, & E. Yoshida (Eds.), *Studies in Modern English: The thirtieth anniversary publication of the Modern English Association* (pp. 21-36). Tokyo: Eihosha.

Tottie, G. (1981). Negation and discourse strategy in spoken and written English. In D. Sankoff & H. Cedergren (Eds.), *Variation omnibus* (pp. 271-284). Edmonton: Linguistic Research.

Tottie, G. (1988). *No*-negation and *not*-negation in spoken and written English. In M. Kytö, O. Ihalainen, & M. Rissanen (Eds.), *Corpus linguistics, hard and soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora* (pp. 245-265). Rodopi.

Varela Pérez, J. R. (2007). Negation of main verb *have*: Evidence of a change in progress in spoken and written British English. *Neuphilologische Mitteilungen, 108*(1), 223-246.

Yadomi, H. (2015). The regulation of the auxiliary *do*: *Do*-less negative declarative sentences in American English from 1800 to the Present Day. *Zephyr, 27*, 44-70.

**Kaltenböck, Gunther**

***University of Graz***

**gunther.kaltenboeck@uni-graz.at**

**Ten Wolde, Elnora**

***University of Graz***

**elnora.ten-wolde@uni-graz.at**

Type of Contribution: Full Paper

# A *just so* story

## On the recent development of the complex subordinator *just so*

*Abstract*

This paper analyzes the phrase *just so* used as a subordinator, as illustrated in (1) and (2) below, which has received very little attention in the literature. We argue that *just so* represents an emerging subordinator of purpose, closely tied to the recent emergence of the pragmatic marker *just so you know*, and discuss possible reasons for this development. The study is based on data from the *Corpus of Contemporary American English* (COCA) and the *Corpus of Historical American English* (COHA), which are analyzed both quantitatively and qualitatively.

In the literature, in the rare cases where the *just so* subordinator is mentioned at all, it is identified as an informal subordinator of condition, as in (2) (Quirk et al., 1985, p. 1090; Kortmann, 1997, pp. 315-316). The data from COHA, however, reveal that there are two distinct uses of the *just so* subordinator, viz. with the meaning of (i) purpose, or (ii) condition, illustrated in (1) and (2) respectively.

(1) *I definitely need more Legos, so we need to have kids* ***just so*** *I can justify the toys.* (COHA, 2004)

(2) "*Faith, Sidony, I don't care how the man arrives,* ***just so*** *he does," Sorcha said impatiently* (COHA, 2006)

The COHA data for the last two centuries, which comprises a total of 776 tokens of *just so* followed by a clause (518 purpose and condition, the rest manner adverbials), show that *just so* as a subordinator is first attested in the 1860s with the two functions condition and purpose emerging simultaneously. From then onwards the subordinator use shows a modest but steady increase (to 3.42 instances pmw in the 2000s).

After an initially parallel development of the purpose and condition uses, the *just so* subordinator develops a clear preference for purpose in the late 20th century at the expense of the condition use, which is rapidly decreasing. This preference is attested in the normalized data as well as in the relative proportion of each semantic type and in the relative change within the set of *so that* subordinators (see Figure 1 for the latter) and can be interpreted as a case of semantic specialization.

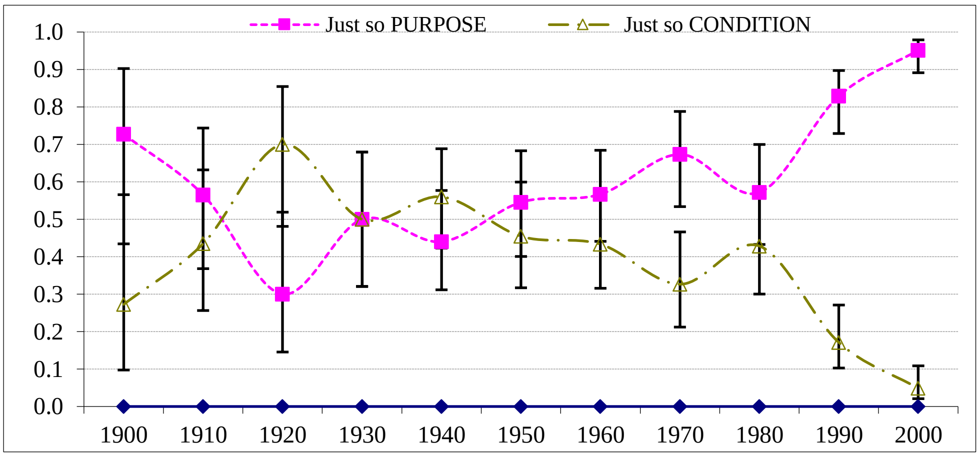


Figure 1: Relative change of purpose and condition subordinators compared to the baseline of all just so subordinators (Wilson confidence intervals for $p < 0.05$; Wallis, 2012).

The steady increase of the purpose *just so* coincides with the emergence of the *just so you know* pragmatic marker, which is first attested in the 1980s in COHA. Although the overall figures are relatively low (e.g. 248 in COCA, 2015-19), the pragmatic marker use shows a clear rise in recent decades and can be taken as further indication of the establishment of *just so* as a purpose subordinator in its own right, i.e. independent of *so (that)*.

In its pragmatic marker use the phrase *just so you know* is positionally mobile, occurring in initial, medial, and final position with regard to a host clause (COCA: 74.9%, 2.4%, 21.4% respectively; 1.3% cases are unclear) and has adopted a range of pragmatic functions. For initial position these include: (i) indicating a topic or focus shift, (ii) indicating an elaboration to a preceding utterance, (iii) expressing emphasis, (iv) preempting a potentially negative implicature. Each of these uses is illustrated in (4)-(7) respectively:

(4) KOTB: You missed the Cutest Baby Contest and we're so bummed because - oh! Come on. STOCKMAN: That's Juice-Juice. KOTB: Is that what you call her? Why? STOCKMAN: That's what I call her. Because of her cheeks. Look at those cheeks. KOTB: Yeah. STOCKMAN: She's got - it's filled with juice. KOTB: Oh! Oh! STOCKMAN: So that's why I call her Juice-Juice. That's my baby. KOTB: **Just so you know**, Shawn's on the road all the time. I didn't realize that you guys were - you were still touring as often as you guy s do. (COCA, 2011)

(5) SPRINGER: […] Thank you. Welcome to the show. Today we're going to make some holiday wishes come true by reuniting our guests with their long-lost family members. OK. Now **just so you know** -- **just so you know**, our guests making the plea don't know that we've found their loved ones, so I'm going to ask everyone not to give it away when -- when we bring our guests out. OK. (COCA, 1996)

(6) HODA-KOTB# […] what a cute video pieces we have that for you. There is a baby elephant - KATHIE-LEE-GIFFORD# That's what I feel like today. HODA-KOTB# - in a kiddie pool. This is in Fort Worth, Texas. Her name is Belle. She was born on July 7th. She weighs three hundred pounds. KATHIE-LEE-GIFFORD# Oh, my gosh. Look at that. HODA-KOTB# Oh, my gosh. Okay. **Just so you know** that is the cutest three hundred-pound baby. KATHIE-LEE-GIFFORD# Oh, my gosh. (COCA, 2013)

(7) Honestly, it's probably weirder to me than it is to you. Look, I can't believe I have to say this, but **just so you know**, there's no way the two of you work. (COCA, 2019)

To explain the development of the *just so* subordinator, it is necessary to look not only at *just so* in isolation but also at the larger network of related subordinator constructions (e.g. Traugott, 2018; Diessel, 2019) and the taxonomic links of *just so* to the formally and functionally related constructions *so* and *so that* (e.g. Verstraete, 2007; Kortmann, 1997, p. 332; Schiffrin, 1987; Quirk et al., 1985, p. 1070). Based on preliminary results for *so that* and *so* in COHA, the following scenario suggests itself. With its increasing semantic specialization, *just so* provides for an informal alternative to *so that*, which lacks semantic precision. *Just so* also offers a useful alternative to the conjunction *so* with its predominant result meaning and inherent mulitifunctionality, which has also been increased by its discourse marker use (Bolden, 2009). *Just so* thus fills the niche of an informal purpose subordinator and in doing so brings greater clarity to the fluctuating semantics of this family of constructions.

### *References*

Bolden, G. B. (2009). Implementing incipient actions: The discourse marker ‘so’ in English conversation. *Journal of Pragmatics, 41*, 974-998.

Diessel, H. (2019). *The grammar network: How language structure is shaped by language use*. Cambridge: Cambridge University Press.

Kortmann, B. (1997). *Adverbial subordination: A typology and history of adverbial subordinators based on European languages*. Berlin: De Gruyter Mouton.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language.* London: Arnold.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.

Traugott, E. (2018). Modeling language change with constructional networks. In P. Bordería, S, Loureda Lamas & Ó. Loureda Lamas (Eds.), *Beyond grammaticalization and discourse markers* (pp. 17-50). Leiden: Brill.

Verstraete, J. (2007). *Rethinking the coordinate-subordinate dichotomy: Interpersonal grammar and the analysis of adverbial clauses in English*. Berlin: De Gruyter**.**

Wallis, S. A. (2012). That vexed problem of choice: Reflections on experimental design and statistics with corpora. London: UCL Survey of English Usage. Retrieved from: www.ucl.ac.uk/english-usage/staff/sean/resources/vexedchoice.pdf

**Kirk, John**[1]

**University of Vienna**

**info@johnmkirk.co.uk**

Type of Contribution: Full Paper

## The digital lexical atlas of Scotland Corpus

New opportunities for researching the vocabulary of Scots

*Abstract*

**Research Questions**

- What are the lexemes of Scots in the *Linguistic Atlas of Scotland* (LAS) (Mather & Speitel, 1975, 1977)?
- In those lexemes, what lexical patterns are present?
- What etymological sources do those lexemes have?
- What does a mapping of those lexemes reveal about regional variation in Scotland?
- How well do lexical and phonological variation corelate regionally within Scotland?
- What insights about the lexical composition of specific areas of Scotland may be derived from the new atlas?
- What does a study of those lexemes reveal about the words used for particular concepts and more generally about lexicalization processes?
- In what ways will the georeferenced lexical data be linkable with georeferenced data from other sources?

[1] With assistance from the other members of the research project team currently comprising Christian Hessle, Ludwig Breuer and Hansi Breuer, and from Malinda Perera for the video production, all gratefully acknowledged)

### Approach

To digitize and map afresh all the present data - including every orthographic variant that wasn’t included in the atlas

- To interpret every orthographic variant as belonging to a lexical type or lexeme and produce new maps - strictly lexical maps
- To create innovations and interactivity in the visual presentation of lexical material
- To interpret the relationship between the physical map and the cartography superimposed on it, and between lexical maps and any other map sharing georeferenced data

### Method

- To undertake a prototypical study (Kirk & Munroe, 1989; Kirk, 1994; Hessle, 2019; cf. Fleischer et al., 2017)
- To construct an online relational database for the storage of the resultant lexemes
- To link the postgreSQL and Django database with appropriate mapping software for storing and editing, processing and accessing
- To display and project the georeferenced data on an interactive online GIS / mapping application, to allow for selecting and displaying any information as desired, so that one could start anywhere in the system and end up anywhere else within the system
- To integrate etymological or cultural or other such information including photographs depicting the lexical references for use and display as required
- Complication: LAS data have old co-ordinates and occasionally very vaguely named localities, whereas GIS/GeoJSON maps work with new co-ordinates.

### Data

- *Linguistic Atlas of Scotland* (LAS) (Mather & Speitel, 1975; 1977)
- Maps: Vol. 1: 90 + Vol. 2: 80 = 170

- <u>Places</u>: Scotland: Vol. 1: 1057, Vol. 2: 569
- <u>Respondents</u>: Scotland: Vol. 1: 1337; Vol. 2: 659; Details of age and sex
- <u>Responses</u>: 1337 respondents x 90 maps + 659 respondents x 80 maps = 173,050; plus multiple responses, minus blank responses<u>Corpus</u>: covers all of Lowland Scots; representative of traditional vocabulary through many lexical fields, e.g. parts of the body, clothes, cooking, domestic architecture, children's games, education, tools, drainage, crops, farm animals, birds, insects, plants

**Prototype Study**

North-Mid Scots - Perthshire, Kinross-shire, Fife & Clackmannanshire (cf. Hessle, 2019; Hessle & Kirk, forthcoming; Kirk & Hessle, forthcoming).

Item 1: 'splinter'

- *LAS-1*: Map 4: *skelb*, *skelf, skelve, spale, spelk, spilk, splice, splinter, spell, stab* and *stob* (x 11 types)
- Data list: *brag, brog, brug, bruggle, drob, jag, jobe, scare, skale, skel, skelb, skelf, skellock, skelp, skelpe, skelve, skelye, sliver, slivver, spale, speel, speld, speldron, spelf, spelk, splinter, stab, steuog, stick, stob, stub, thorn* (x 32 orthographic types).
- Lexemized: *scare* 1, *skale* 1, *skelb* 134, *skelf* 64, *sliver* 5, spale 45, *splinter* 18, *stab* 14, *stug* 1, *stick* 1, *thorn* 1; plus Onomatopoeic 7: *brog*, *drab*, *jag*,
- Items for discussion: *skelb/skelp* vs. *skelf/skelve*, which raise an etymological issue. *Skelb*/*skelp* are from Gaelic *sgealb* (cognate with Irish *scealb*, whereas *skelf* (*skel*, *skelve* and *skelye*) are of Dutch origin (*schelf* CSD, p. 629). The respondents giving both were predominantly from the southern coast of Fife, where Gaelic survived until well after the Middle Ages and where there was considerable Dutch settlement in the twelfth century. We interpret these as separate words, as two separate lexical survivals, not as

pronunciation variants, unlike *skelp/skelb*, or *skelf/skelve*, which were not given by the same respondent.

Item 2: ‘the youngest of a brood’

- *LAS*-1: Map 65: *creet, cricklet, crit, crowl, cruit, crut, dorneedie/do(o)rneed, draidlock, dronie, jory, ricklin(g), rig(gie), scoor da buggie, sharger, tailag, titlin(g), weirdie/wairdie, wreckin(g), youngest* (19 variants)
- Data list: *anthony, baby, bairn, benjamin, benjie, broodie, cheeper, chick, chicken, crit, dorneed, draidlock, draidluck, dredlock, dreidlich, drich, dridlock, droich, droichie, end o' lachter, last of the lachter, reg, rig, riggie, riglin, runt, scrunt, shake of the bag, shakin’ o’ the pokie, shakins o’ the poke, shott, stirk, tail end, titlin, voollie voorin, wairdie, wardie, weardie, wee tot, wee wairdie, weirdie, wreckling, wuirdie, youngest, youngest of cleckin, young one* (49 variants, comprising 123 tokens)
- Lexemized (discussed in Kirk & Hessle, forthcoming):
- denotans: *brood(ie)* 1, *crit* 6, *dorneed* 1, *draidlock* 5, *dreddle* 1, *droich* 4, *rig* 38, *runt* 2, *shott* 1, *weirdie* 42, personal names 5,
- non-denotans: baby names 2, onomatopoeic names 3, *chick(en)* 2, *end/last o lachter* 2, *(shake) o the bag/poke* 3, *stirk* 1, *voolie voorin* 1, *warrock* 1, *wee/young X*
- Items for discussion: supernatural names, sacred/religious names

## Concluding Remarks

By undertaking a strictly lexemic analysis of the variants for each item listed in the *LAS*, our analysis has taken us far beyond the words of the pilot study and Scotland to see the shared systems of religious and folk belief which cut across and transcend languages and nations. In tracing these lexical links there’s a wealth of detective-like enquiry to be done, to be richly facilitated by the digitization of the data and the new interactive possibilities for connections and mappings, for searching and display, for fresh interpretation and description.

Our research questions are being answered: through lexemization; through distinguishing denotans lexemes from non-denotans lexemes, the latter comprising descriptions, metaphorizations, cultural associations as well as semantic transfers; through the visualization of linguistic information through cartography and distinguishing the georeferenced cartography from the physical map.

We are providing a new dynamic platform for new insights not just about lexical variation in Scotland but a resource of georeferenced data more widely utilizable by ethnographers, culture scientists and many other researchers.

## *References*

Fleischer, Jürg, Lenz, A. N., & Weiß, H. (2017). SyHD-atlas. Konzipiert von Ludwig M. Breuer. Unter Mitarbeit von K. Kuhmichel, S. Leser-Cronau, J. Schwalm, & T. Strobel. Marburg/Wien/Frankfurt.

Hessle, C. (2019). *Towards a digital version of the linguistic atlas of Scotland*. Unpublished BA dissertation, University of Vienna.

Hessle, C., & Kirk, J. (forthcoming). Digitising collections of historical linguistic data: The example of The Linguistic Atlas of Scotland. *Journal of Data Mining and Digital Humanities.*

Johnson, P. (1997). Regional variation. In J. Charles (Ed.) *The Edinburgh history of the Scots language* (pp. 443-513). Edinburgh: Edinburgh University Press.

Kirk, J. (1994). East Central Scots: A computerised mapping package. In A. S. Fenton & D. A. MacDonald (Eds.) *Studies in Gaelic and Scots* (pp. 48-68). Edinburgh: Canongate Academic.

Kirk, J., & Hessle, C. (forthcoming). Towards a digital lexical atlas of Scotland. *Scottish Language*.

Kirk, J., & Munroe, G. (1989). Dialectometry: From corpus lists to analyzed maps. Paper presented at ICAME10, Bergen.

Kirk, J., Hessle, C., Breuer, L., & Breuer, H. C. (forthcoming). *The digitised lexical atlas of Scotland*. University of Vienna.

Mather, J. Y., & Speitel, H. H. (Eds.). (1975). *The linguistic atlas of Scotland. Scots section.* Vol. 1. London: Croom Helm.

Mather, J. Y., & Speitel, H. H. (Eds.). (1977). *The linguistic atlas of Scotland. Scots section.* Vol. 2. London: Croom Helm.

Tulloch, G. (1997). Lexis. In J. Charles (Ed.) *The Edinburgh history of the Scots language* (pp. 378-432). Edinburgh: Edinburgh University Press.

**Klavan, Jane**

***University of Tartu***

**jane.klavan@gmail.com**

**Roomäe, Kärt**

***University of Tartu***

**kart.roomae@gmail.com**

**Savchenko, Denys**

***University of Tartu***

**denys.savchenko@ut.ee**

Type of Contribution: Oral Presentation

## A corpus study of intensifiers in Estonian learners' spoken English

*Abstract*

Our study has two broad aims: to describe the process and the current status of compiling the Estonian subcorpus of LINDSEI; and to give a preliminary overview of the most frequent intensifiers in the spoken language of Estonian EFL learners using LINDSEI-EST as our target corpus.

### LINDSEI-EST: Compilation and current status (work-in-progress)

We are currently compiling the Estonian subcorpus of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) at the English Department of the University of Tartu. The LINDSEI database was launched in 1995, and it contains oral data produced by advanced EFL learners with different language backgrounds. As of 2019, 13 projects have been completed, 9 are in progress. (Université catholique de Louvain, n.d.) The principal goal of the LINDSEI project is to collect comparable data in order to analyze the over-

and underuse of linguistic patterns from a cross-linguistic perspective that allows to determine whether these trends are language specific or universal (Pravec, 2002, p. 83). The comparability criterion also explains why the compilers of the corpus have given specific instructions with regard to how the interviews and transcription should be conducted.

The interviews are structured around three tasks: three conversation topics (choice between a valuable experience, a country visit that left an impression, or a memorable film/play they have seen), free discussion and a picture description task. In the last task, the interviewees have to tell a story based on four pictures depicting an artist who is painting a portrait of a woman. They are not allowed to take any notes. The interview takes approximately 15 minutes in total. In addition to obtaining informed consent, we are collecting metadata according to the LINDSEI guidelines about participants' age, gender, nationality, country, native language, education background, stays in an English-speaking country, and the foreign languages they speak.

At its current stage, the LINDSEI-EST corpus consists of 17 interviews (about 227 minutes of speech; 28,495 words of transcribed text). The 17 interviews were recorded in 2018 using Tascam DR-05, a 24-bit/96kHz Digital Recorder. All of the interviewees (12 female, 5 male; average age 23 years) were native speakers of Estonian. They were third or fourth year students of English language and literature at the University of Tartu.

## Intensifiers in Estonian advanced learners' spoken English: a pilot study

Intensifiers are concerned with the semantic category of degree and are commonly divided into amplifiers and downtoners (Quirk et al., 1985). Amplifiers scale quality upwards and downtoners scale quality downwards from the assumed norm. Amplifiers are divided into maximizers and boosters. Maximizers (e.g. *completely*) express the upper point on the scale and boosters (e.g. *very*) express a high degree or point of the scale. Ito & Tagliamonte (2003) stress that variation in the use of intensifiers is a strong indicator of shifting norms and practices; studying such a linguistic phenomenon can, therefore, make an important contribution to discovering the current trends in learner language.

In this study we focus on intensifiers that function as amplifiers and that modify adjectival heads only (cf. Ito & Tagliamonte, 2003, p. 258; Xiao & Tao, 2007). According to Bäcklund (1973, p. 279), the majority of intensifiers are used to pre-modify adjectives. This approach will allow us to collect data that is comparable with the previous findings regarding the use of intensifiers in the spoken language of native speakers, and to draw conclusions regarding the current trends in learner language. For native language, Palacios Martínez & Núñez Pertejo (2012) show that *really* and *so* are the most frequent intensifiers among British teenagers. Ito & Tagliamonte (2003) examined the use of intensifiers in a corpus of the native population of the city of York, England; they show that the most frequent intensifiers are *very* and *really*.

As for learner language, the results are inconsistent. For written language, a study by Granger (1998) found a statistically significant underuse of amplifiers in the learner corpus, both in the number of tokens and types. Lorenz (1999) and Recski (2004), however, report EFL writer's overuse of intensifiers. Lorenz (1999) attributes this to the learners' tendency towards 'information overcharge' - one of the main stylistic weaknesses of non-native writing. For spoken language, a study by Pérez-Paredes (2010) seems to confirm that amplifiers are not part of the active spoken lexical repertoire of learners when compared to native speakers. These inconsistencies are likely due to the different ways researchers have operationalized "intensifiers", the L1 of learners, and the mode (written *vs* spoken).

## Data extraction and data analysis

The aim of the pilot study is to give a preliminary overview of the distribution of intensifiers in LINDSEI-EST. The following procedure was used:

(1) Data cleaning: filled pauses and backchanneling were removed; the interviewer's productions were eliminated.

(2) A utomatic POS-tagging: we used spaCy (https://spacy.io/) as the POS-tagging tool. The tool offers reasonably accurate tags when annotating spoken data despite various mistakes connected to the nature of spoken language.

(3) Data extraction: we used Python together with spaCy for extracting grammatical items and the NLTK for concordance lines. After

running several tests, we used two patterns: (Pattern 1) optional dependency tag 'advmod' (adjective modifier), compulsory tag 'advmod' and POS-tag 'ADJ' (adjective). This pattern was chosen in order to retrieve both adverb-adjective collocations as well as adverb-adverb-adjective combinations. (Pattern 2) dependency tag 'advmod' (adjective modifier) and POS-tag 'VERB'. This pattern was added to extract examples like *really dumbed down sentences* as the POS-tagger assigned the tag 'VERB' to the adjective *dumbed-down*. (4) Manual data annotation: the extracted data were manually annotated for the type of intensifiers (maximizer, booster, downtoner, negation, other, ambiguous). All the contexts that do not allow intensification, in which intensifiers functioned as downtoners or appeared under negation, were excluded from the subsequent analysis.

**Preliminary results**

The data extraction process produced a total of 362 ADV + ADJ combinations in the LINDSEI-EST data. For amplifiers, we found a total of 183 boosters and 18 maximizers. The data included 10 different types of boosters (*actually, how, particularly, properly, really, so, that, too, truly, very*) and 7 different types of maximizers (*completely, entirely, extremely, just, quite, super, totally*). There are considerable differences in the frequency of intensifiers used by individual learners, ranging from 13 uses per 1,000 words to only 4 uses per 1,000 words. Our results confirm some of the previous findings - the three most frequently used intensifiers are *very* (65x)*, really* (63x)*, so* (37x). As for future work, we want to extend the notion of "intensification", compare the use of ADV + ADJ intensifiers to native data and other L1 learner data. Most importantly, we need to collect more Estonian EFL data for both spoken and written language.

## *References*

Bäcklund, U. (1973). *The collocation of adverbs of degree in English*. Acta Universitatis Upsaliensis.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and lexical phrases. In A.P. Cowie (Ed.), *Phraseology:*

*Theory, analysis and applications* (pp. 145-160). Oxford: Clarendon Press.

Ito, R., & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society*, *32*, 257-79.

Lorenz, G.R. (1999). Adjective intensification - learners versus native speakers: A corpus study of argumentative writing. Amsterdam: Rodopi. *Language and Computers: Studies in Practical Linguistics, 27*.

Pérez-Paredes, P. (2010). The death of the adverb revisited: Attested uses of adverbs in native and non-native comparable corpora of spoken English. In Moreno Jaén, M., Serrano Valverde, F. & M. Calzada Pérez (Eds.), *Exploring new paths in language pedagogy: Lexis and corpus-based language teaching* (pp. 157-172). London: Equinox.

Martínez, I.M.P., & Pertejo, P.N. (2012). He's absolutely massive. It's a super day. Madonna, she is a wicked singer. Youth language and intensification: a corpus-based study. *Text & Talk*, *32*(6), 773-796.

Pravec, N.A. (2002). Survey of learner corpora. *ICAME Journal 26*, 81-114.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language.* London: Longman.

Recski, L.J. (2004). "It's really ultimately very cruel": contrasting English intensifier collocations across EFL writing and academic spoken discourse. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 20*(2), 211-234.

Université catholique de Louvain. (n.d.). Centre for English Corpus Linguistics. *LINDSEI*. Retrieved from https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html

Xiao, R., & Tao, H. (2007). A corpus-based sociolinguistic study of amplifiers in British English. *Sociolinguistic studies, 1*(2), 241-273.

## Ireland Kuiper, Katie

***University of Georgia***

**katherine.kuiper25@uga.edu**

Type of Contribution: Work-in-Progress Report

# The e-cigarette dilemma

*Abstract*

E-cigarettes are currently the most common tobacco products in use by youth, “driven in large part by marketing and advertising by e-cigarette companies” (Walley et al., 2019). The following presentation discusses an ongoing pilot study concerning views surrounding e-cigarettes in both US and UK press. Previous studies have demonstrated the utility of corpus research concerning health communication (Hunt & Harvey, 2015), press views of health topics (Hannaford, 2017, 2019; Potts & Semino, 2017; Semino et al., 2015), tobacco advertising (Kretzschmar et al., 2004), and rhetorical strategies used by the press (Baker et al., 2013; McEnery & Baker, 2017). Current research questions include the following: How have press views concerning vaping and e-cigarettes changed over time in both the US and the UK? What rhetorical and linguistic strategies are used over time concerning e-cigarettes and e-cigarette users, and what methods best address these strategies?

This ongoing study utilizes principled sampling to create a representative corpus for news sources in both the United States and the United Kingdom, to first establish and understand the regularly used rhetorical, persuasive, and linguistic strategies in the press, with regards to vaping usage in both places, to identify areas in which manipulation may have occurred, and to understand the ongoing discourse over time with regards to this particular health-related technology. Articles were obtained from US news sources including *The New York Times*, *USA Today*, *The Washington Post*, *The Wall Street Journal, The Atlanta Journal Constitution, The New York Post, The Baltimore Sun, The Arizona Republic, The Boston Globe, The Chicago Tribune, Detroit Free Press, The Journal Record, The Los Angeles Times, The Philadelphia Inquirer,*

*Newsday, Portland Press Herald, Salt Lake Tribune, South Florida Sun-Sentinel, St. Louis Post-Dispatch, Star Tribune, Texas Tribune,* and *NPR*. The UK corpus includes articles from *The Guardian*, *The Telegraph, The Financial Times*, *The Belfast Telegraph, The Daily Mail, Edinburgh Evening News, Mail on Sunday, Manchester Evening News, Sunday Mirror, The Birmingham Post, The Daily Mirror, The Financial Times, The Herald, London Evening Standard, Wales on Sunday, The Times,* and *The Sun*. Both corpora are being added to continually, but currently the UK Corpus is at approximately 1.2 million words, and the US Corpus is at just over 2.2 million words. Each corpus was created with data obtained courtesy of the UGA Library Databases: Proquest and Gale OneFile News. All articles were selected from the following sources if they contain the term *e-cigarette(s*) at least once in the text. The raw text files and titles were then added to separate documents, and the metadata for all relevant article information, including article source, date, year, type, and geographic location, was collected in a separate spreadsheet. This process is still ongoing for both corpora, and they contain data from 2009, through the present day (April 2020). In order to analyze the data within in corpus, I use the command-line version of CQP (Evert et al., 2017), R packages polmineR() and RcppCWB() (Blätte, 2019; Blätte et al., 2019), mallet topic analysis (McCallum, 2002; Weingart, n.d.), and traditional corpus methods including keyword, collocational, and frequency analysis.

The presentation discusses the results of both collocational analysis, mallet topic analysis, as well as future plans for research. Current findings reveal differences in choices of emphasis by press outlets in the US and UK corpora, as well as change over time, reflecting current conflicted views and public health concerns with e-cigarette use (Walley et al., 2019). The following group of charts display the results for collocational analysis for terms modifying *e-cigarettes* over time in the UK, from 2014, 2019, and 2020.

As shown below in Figure 1, the top collocates in the UK change over time, so that the term *use* is the most frequent collocate in 2014, but the most frequent in 2019 press involves terms like *ban, flavored,* and *Trump.* In 2020, *believe* is one of the most frequent collocates, which could be associated with readers and authors beliefs about the efficacy of e-cigarettes in general; *ban* is also a top-most collocate during this year. In addition to this example, I subset collocational analysis by different metadata, including year, geographical location, and type of source.

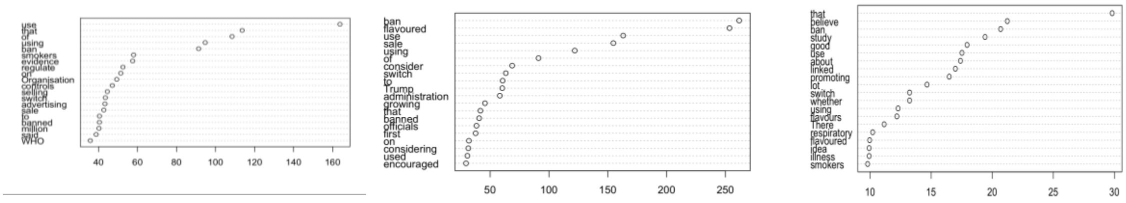


Figure 1: Collocates modifying *E-Cigarettes* UK.

These are discussed more thoroughly in the presentation. Mallet topic analysis output matches that of relevant collocates for the terms *e-cigarette(s)* in both the US and the UK, and brings up interesting topics that are found throughout both corpora. Topic groups for the US corpus include topic groups of tokens associated with research, politics, the e-cigarette industry, schools, and users. Topic groups for the UK corpus include those groups as well as tokens associated with regulation, and the movie industry. Both corpora include terms associated with relevant politicians (namely Trump) and places where e-cigarette use is of interest or concern.

Because this is a pilot study, I am continuing to analyze the dataset and gather more data. I plan to continue with collocational and keyword analysis, as well as consider other ways that linguistic and rhetorical strategies change by location and political views in both the US and the UK. This research is of great interest and importance not only for linguists, but also for anyone concerned with public health, the spread of information, and the way that press sources communicate about health. This research underscores important questions concerning how public health information and data is shared, and there is much more work to be done in this arena.

## *References*

Atkins, S., & Harvey, K. (2010). How to use corpus linguistics in the study of health communication. In A. O'Keefe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 605-619). New York: Routledge.

Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press.* Cambridge: Cambridge University Press.

Bowen, M. (2019). Stigma: A linguistic analysis of personality disorder in the UK popular press, 2008-2017. *Journal of Psychiatric and Mental Health Nursing, 26*(7-8), 244-253.

Blätte, A., Desgraupes, B., Louiseau, S., Christ, O., Schulze, B. M., Evert, S., & Fitschen, A. (2019). RccpCWB package, v 0.2.8.

Blätte, A., & Leonhardt, C. (2019). PolmineR package, v 0.8.0.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139-173.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide.* Cambridge: Cambridge University Press.

Casaa.org. (2012-2019). Historical timeline of electronic cigarettes. Date accessed: 5 Oct 2019.

Evert, S., & the CWB Development Team. (2017). The IMS Open Corpus Workbench (CWB) Corpus Encoding Tutorial.

Gries, S. (2009). What is corpus linguistics? *Language and Linguistics Compass, 3*(5), 1225-1241.

Hannaford, E. (2017). The press and the public attitudes on mental health: A corpus linguistic analysis of UK newspaper coverage of mental illness (1994-2014), compared with the UK national attitudes to mental illness survey. MPhil Thesis, University of Glasgow.

Hardie, A. (2012). CQP Web: Combining power, flexibility, and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*(3), 380-409.

Hunt, D., & Harvey, K. (2015). Health communication and corpus linguistics: Using corpus tools to analyse eating disorder discourse online. In P. Baker & T. McEnery (Eds.), *Corpora and Discourse Studies* (pp 134-154). London: Palgrave Macmillan.

Jasal, R., & Nerlich, B. (2017). Polarised press reporting about HIV prevention: Social representations of pre-exposure prophylaxis in the UK press. *Health,* 21(50), 478-497.

Kretzschmar, W., Darwin, C., Brown, C., Rubin, D., & Biber, D. (2004). Looking for the smoking gun: Principled sampling in creating the tobacco industry documents corpus. *Journal of English Linguistics, 32*(1), 31-47.

Mautner, G. (2008). Analyzing newspapers, magazines and other print media. In R. Wodak & M. Krzyżanowski (Eds.), *Qualitative discourse analysis in the social sciences* (pp. 30-53). Basingstoke: Palgrave Macmillan.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

McEnery, T., & Baker, H. (2017). *Corpus linguistics and 17th century prostitution*. London: Bloomsbury Publishing.

McEnery, A., Brezina, V., & Baker, H. (2019). Usage fluctuation analysis: A new way of analyzing shifts in historical discourse. *International Journal of Corpus Linguistics*, *24*(4), 413-444.

Patra, B.G., Ghosh, N., Das, D., & Bandyopadhijay, S. (2015). Identifying temporal information and tracking sentiment analysis in cancer patients interviews. *Computational linguistics and intelligent text processing: 18th CICLing Proceedings Part 2* (pp. 180-188). Springer Publishing.

Potts, A., & Semino, E. (2017). Healthcare professionals' online use of violence metaphors for care at the end of life in the US: A corpus-based comparison with the UK. *Corpora, 12*(1), 55-84.

Semino, E., Demnjén, Z., Demmen, J., Koller, V., Payne, S., Hardie, A., & Rayson, P. (2015). The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: A mixed methods study. *BMJ Supportive and Palliative Care, 7*(1), 60-66.

Sinclair, J. (2004). *Trust the text: Language corpus and discourse*. London: Routledge.

Walley, S. C., Wilson, K. M., Winickoff, J. P., & Groner, J. (2019). A public health crisis: Electronic cigarettes, vape, and JUUL. *Pediatrics, 143*(6), e20182741.

Weingart, S. n.d. Topic modeling for humanists: A guided tour. Retrieved from: scottbott.net.

Whitley, R., & Wang, J. (2017). Good news? A longitudinal analysis of newspaper portrayals of mental illness in Canada 2005-2015. *The Canadian Journal of Pyschiatry, 62*(4), 278-285.

**Laitinen, Mikko**

***University of Eastern Finland / Linnaeus University***

**mikko.laitinen@uef.fi**

**Fatemi, Masoud**

***Linnaeus University / University of Eastern Finland***

**fatemi@cs.uef.fi**

Type of Contribution: Full Paper

## Working with complex metadata

Computational approaches to digital social networks

*Abstract*

This presentation uses complex corpus metadata and focuses on social networks by combining sociolinguistics and data mining. Building on the work by Granovetter (1973), social network theory in sociolinguistics assumes that individuals form personal communities that provide a meaningful framework for them in their daily life. An individual's social network is the sum of relationships contracted with others, and networks may be characterized by ties of varying strength. If ties are strong and multiplex, the network is dense, and individuals are linked through close ties (such as friends). Conversely, ties can be weak in which case individuals are predominantly linked through occasional and insignificant ties (such as acquaintances), and the network is loosely knit. Most importantly, networks play a substantial role in language maintenance and change. Empirical evidence shows that loose-knit networks promote innovation diffusion, whereas dense multiplex networks resist change (Milroy & Milroy, 1978; Milroy L., 1987). The underlying reason for the weakness of strong ties in transmitting innovation is the fear of losing one's social standing in a network. Adopting new ideas is socially risky, and we do not want to "rock the boat" in dense social structures.

While the social network theory is influential in sociolinguistics, it is based on very small data, and most studies have focused on ego networks obtained using ethnographic observations. According to Milroy & Milroy (1992, p. 5), this “effectively limits the field of study, generally to something between 30 and 50 individuals”. Moreover, it has been suggested that the quantitative variable of a network “cannot be easily operationalized in situations where the population is socially and/or geographically mobile” (Milroy, 1992, p. 177). In this presentation, we report a study that concentrates on networks that are larger than small networks of only a few dozen of individuals. This has been done because evidence from social anthropology suggests that average human networks are substantially larger (>150 nodes) (Dunbar, 1992; McCarthy et al., 2001). Prior empirical work in sociolinguistics has, therefore, covered only a limited section of possible network sizes.

We have two research questions. First, we test the extent to which social media data from Twitter and computational methods could be utilized to operationalize network ties of highly mobile individuals in very large datasets. Second, we specifically concentrate on the effect of network size on the validity of the theory. We investigate if the fear of losing one’s social standing by rocking the boat disappears in large strong-tie networks. To respond to these questions, we discuss a computational method that can take up large and messy social media data and render these data usable for analyzing networks, thus expanding the empirical basis substantially.

To answer the research questions, we use mutual interaction metadata from freely available Twitter data from Sweden and Finland. These mutual interaction data are subjected to algorithms to assign labels of weak or strong networks to the accounts. The algorithms are mainly from the graph theory and the set theory, and some of them have been developed by us. First, we use betweenness measures that identify nodes that act as brokers between communities and are used to detect the density of how people are connected to each other in a network. These include betweenness centrality, which is a measure based on finding the shortest path between nodes (Freeman, 1977) and closeness centrality (Perez & Germon, 2016). Second, we also use Jaccard Similarity Coefficient, a symmetric measure that calculates the similarity between two sets, and it is used to measure the similarity between accounts in terms of the number of common followers/friends. The assumption is that the share of common friends/followers is higher in a strong-tie network than in weak-tie settings. In addition, we assign weights to each account in the network to measure frequency of communication between network nodes and employ a method which we call disjointness. This last method enables us to estimate how well the nodes in a network are connected if the ego node were to be removed. The network labels are therefore multidimensional.

As for the dependent variables, we employ items that are frequent enough in our data. First, we use the share of English messages in each network. We rely on automatically-assigned language labels in the metadata. Second, we use a mixture of linguistic features available in the tweet text. These features consist of contracted forms (*won't*, *'ll*, *I'm* etc.), and NEED *to* used as a semi-modal auxiliary. These features are qualitatively different as the contracted forms index colloquial, spoken-like use, while *NEED to* is currently undergoing change in English and is highly pervasive in ELF use in the Nordic region.

Our results show that the method enables us to extract large digital networks with differing qualities (both weak and strong-ties). These networks cannot only be visually confirmed (Figure 1 below), but they are also quantitatively verified.

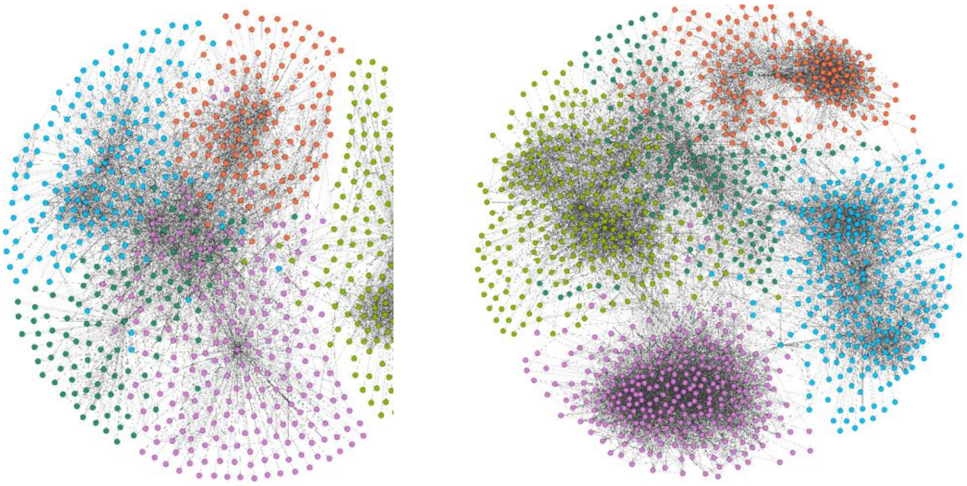


Figure 1: Visualizing weak (left) and strong-tie (right) networks extracted through the algorithms.

As we show in our spoken presentation, the quantitative patterns in the data are clear. When we investigate large networks whose sizes are closely resemble average human networks, we can observe identical patterns. Our results show no distinction between large weak-tie and strong-tie networks, which suggests that the differences observed in small ethnographic studies level out when the network size becomes sufficiently large. These observations not only support the findings in our pilot study (Laitinen et al., 2017), but they also introduce ways of measuring the digital networks of mobile individuals in the social media.

## *References*

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution, 22*(6), 469-493.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40, 35-41.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380.

Laitinen, M. (2016). Ongoing changes in English modals: On the developments in ELF. In O. Timofeeva, S. Chevalier, A. Gardner & A. Honkapohja (Eds.), *New approaches in English linguistics: Building bridges* (pp. 175-196). Amsterdam: John Benjamins.

Laitinen, M., Lundberg, J., Levin, M., & Lakaw, A. (2017). Revisiting weak ties: using present-day social media data in variationist studies. In T. Säily, M. Palander-Collin, A. Nurmi & A. Auer (Eds.), *Exploring future paths for historical sociolinguistics* (pp. 303-325). Amsterdam: John Benjamins.

McCarty, C., Killworth, P., Bernard, H. R., Johnsen, E., & Shelley, G. (2005). Comparing two methods for estimating network size. *Human Organization, 60*(1), 28-39.

Milroy, J. (1992). *Linguistic variation and change*. Oxford: Blackwell.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics, 21*, 339-384.

Milroy, J., & Milroy, L. (1978). Belfast: change and variation in an urban vernacular. In P. Trudgill (Ed.), *Sociolinguistic patterns in British English* (pp. 19-36). London: Edward Arnold.

Milroy, L., & Milroy, J. (1992). Social network and social class: Toward an integrated sociolinguistic model. *Language in Society, 21*, 1-26.

Milroy, L. (1987). *Language change and social networks*. 2nd edition. Oxford: Blackwell.

Perez, C., & Germon, R. (2016). Graph creation and analysis for linking actors: Application to social data. In R. Layton & P. A. Watters (Ed.), *Automating Open Source Intelligence* (pp. 103-129). Waltham: Elsevier.

**Landmann, Julia**

***University of Heidelberg***

**evajulia.landmann@as.uni-heidelberg.de**

Type of Contribution: Full Paper

# The pragmatic-contextual use of French, Spanish, German and Yiddish borrowings in English since the nineteenth century

## A historical and socio-cultural analysis

*Abstract*

This paper concentrates on the pragmatic-contextual usage of several thousand borrowings that have been taken over from French, Spanish, German and Yiddish into English since 1801. The words under scrutiny were retrieved from the *Oxford English Dictionary Online*. On the basis of the linguistic evidence available in the *OED* and in corpora (e.g. the *Corpus of Historical American English*), the pragmatic-contextual usage of the various borrowings was investigated from a historical perspective. The *COHA* currently constitutes the largest structured historical corpus of English, offering essential insights into linguistic change and variation. The number of usage examples from various genres is nearly the same for each decade, which makes it possible to compare the usage of lexical items in the diversity of genres (encompassing spoken language) over time.

An essential aim of the present paper is to assume a historical perspective on the use of the borrowings under review in informal language. Besides the *COHA*, further corpora such as the *Spoken British National Corpus 2014* and the *Corpus of American Soap Operas* will be consulted as they reflect informal usage. The *Spoken BNC 2014* consists of 11.5 million words. It comprises 1251 authentic everyday conversations by 627 native British speakers. The *Soap Corpus* documents very informal American English usage. It contains 100 million words of texts retrieved from American soap operas which were produced at beginning of the 21st century.

The linguistic data will be investigated in relation to the socio-cultural or historical contexts relevant for the adoption of the foreign-derived vocabulary into English. It will be essential to determine what is perceived to be ‘French’, ‘Spanish’, ‘German’ or ‘Yiddish’ when comparing the contextual use of the various borrowings. This also raises the question of whether the linguistic documentary evidence reveals culture-specific attitudes of the different speakers.

The use of German-derived words reflecting German-American identity is not as common as the equivalent phenomenon in Yiddish, where lexical items of Yiddish origin are consciously used to create “ethnolinguistic repertoires” (in the sense of Benor, 2010). According to Benor, the term denotes “a fluid set of linguistic resources that members of an ethnic group may use variably as they index their ethnic identities” (2010, p. 160) and “the arsenal of distinctive linguistic features available to members of a given group” (2010, p. 162). Benor (2010, p. 160) outlines that the typical linguistic peculiarities encompass “system-level morphosyntactic, phonological, and prosodic features, as well as sporadic lexical and discourse features.” Benor (2010, p. 164) points out that

> Jews who are part of more traditional denominations, especially Orthodoxy, and participate more in religious life are more likely to use Hebrew and Yiddish loanwords and Yiddish constructions, and these linguistic resources enable them to indicate to others not only that they are Jewish but also that they are a certain type of Jew.

As will become clear, several lexical items of Yiddish provenance are perceived as cultural indicators by (American) Jews, in order to convey an authentic portrayal of Jewish culture and to reveal their ethnic identity. There are some words of Yiddish origin which specify concepts characteristic of Jewish culture. They can therefore be assigned to the group of borrowings which can be used to create ethnolinguistic repertoires. The Yiddish-derived word *shul* serves as an example. It occurs in the meaning of ‘synagogue’ in Jewish usage. It was ultimately derived from the German *Schule* ‘school’. Another example is *sheitel*, a variety of wig typically worn by Orthodox Jewish women who are married. Its German equivalent *scheitel* has a different meaning. It refers to the ‘crown of the head’ in German, and it does not serve as an ethnic or a cultural clue in German usage.

There are also some borrowings which have to be analyzed against a given historical background in which they entered the English language. An example is the word *diktat*. From the *OED3* it emerges that it was assumed from German into English in 1922. The early *OED3* usage examples reveal that *diktat* was initially used with reference to the Treaty of Versailles, which marked the end of the First World War in accordance with international legislation:

- "1922 Crown Prince Wilhelm *Memoirs* iii. 90 In the June days just gone by, came the news that the Versailles 'Diktat' had been signed."
- "1931 *Internat. Affairs* **10** 207 The Treaty of Versailles could not be regarded as anything but a '*Diktat*' put on Germany by force."

Some years after its earliest documented usage in English (i.e. in 1941), *diktat* broadened in meaning: it assumed a more general sense, designating any decree or rule. *Diktat* is thus no longer restricted to a political context in English, as is corroborated by two usage examples found in the *OED3*:

- "1968 C. James *Young Lives at Stake* iv. 89 One can have priorities about what one suggests as content of school curricula, but essentialist fiats or unilateral diktats for the curriculum can no longer be issued."
- "2000 *U.S. News & World Rep.* 22 May 54/2 Even the color of the walls in chip plants around the world is specified by company diktat."

As will be seen, there is a variety of historical and socio-cultural contexts that are significant for the analysis of the borrowings undertaken in this study. The present paper will describe several case studies that point to connections of linguistic aspects and socio-cultural attitudes which have been identified in the overall investigation.

## *References*

Benor, S. B. (2010). Ethnolinguistic repertoire: Shifting the analytic focus in language and ethnicity. *Journal of Sociolinguistics, 14*(2), 159-183.

Durkin, P. (2014). *Borrowed words: A history of loanwords in English*. Oxford: Oxford University Press.

Schultz, J. (2012). *Twentieth-century borrowings from French to English: Their reception and development*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Schultz, J. (2016). *Twentieth century borrowings from German to English: Their semantic integration and contextual usage*. Duisburger Arbeiten zur Sprach- und Kulturwissenschaft/Duisburg Papers on Research in Language and Culture. Frankfurt: Lang.

Schultz, J. (2018). *The influence of Spanish on the English language since 1801: A lexical investigation*. Newcastle upon Tyne: Cambridge Scholars Publishing.

**Larsson, Tove**

***Uppsala University***

**tove.larsson@engelska.uu.se**

**Egbert, Jesse**

***Northern Arizona University***

**jesse.egbert@nau.edu**

**Biber, Douglas**

***Northern Arizona University***

**douglas.biber@nau.edu**

Type of Contribution: Full Paper

# Do corpus linguists focus on statistics at the expense of linguistic description?

A ten-year perspective

*Abstract*

Over the past few decades, many subfields of linguistics have seen a steady increase in the use of advanced statistical methods (see, e.g., Gass et al., 2020). However, to date, observations about the trend of increasing quantitative sophistication and its effects on Corpus Linguistics have been based on anecdotal evidence, and few steps have been taken towards investigating them empirically. The present study aims to carry out a systematic analysis of the relative focus on statistical analysis versus linguistic description in 47 articles published in the four major corpus linguistics journals (*International Journal of Corpus Linguistics*, *Corpus Linguistics and Linguistic Theory*, *ICAME Journal* and *Corpora*) in 2009 and 2019.

While meta-analyses examining research trends in the field have been carried out in subfields of corpus linguistics (e.g., Plonsky & Paquot, 2017), little

attention has been paid to the field at large. In the present study, we analyzed the results and discussion sections of research articles, coding for type and number of statistical tests employed, number of words, tables and graphs devoted to statistical reporting vs. linguistic analysis, and number of text excerpts.

*Statistical reporting* is defined as prose devoted to the reporting of findings of quantitative or statistical analysis; *linguistic description* is defined as prose devoted to the reporting of findings of qualitative linguistic research or the interpretation of findings (either qualitative or quantitative). We defined *statistical technique* as any technique that involves a formula, meaning that descriptive statistics such as means are included along with more advanced techniques such as *Confirmatory Factor Analysis*. We included all 47 articles from 2009 and 2019 that were based on empirical analyses of language; we excluded studies that aimed to introduce a new corpus or method. The following research questions guided the analysis:

- What is the proportional distribution of linguistic analysis vs. statistical reporting (operationalized by number of words, tables, figures and examples devoted to each of these) and what statistical tests are employed?
- Has there been a discernable change in these variables between 2009 and 2019?
- What can this tell us about the direction of corpus linguistics as a field?

At first glance, the results show a fairly even distribution of words devoted to linguistic analysis vs. statistical reporting (47.6% vs. 52.4%) in the results, discussion and conclusion sections. However, with time added as a factor, a clear shift became evident: the proportion of prose devoted to statistical reporting increased significantly from 37.2% in 2009 to 62.6% in 2019 ($p$ = 0.000025, $d$ = -1.43). The boxplots in Figure 1 display the interquartile range (the box), the median (the horizontal black line inside the box) and the mean (the red plus sign). The individual data points are marked by black dots.

As can be seen, while 75% of the articles exhibited a preference for statistical reporting over linguistic description in 2009, the exact opposite is true in 2019.

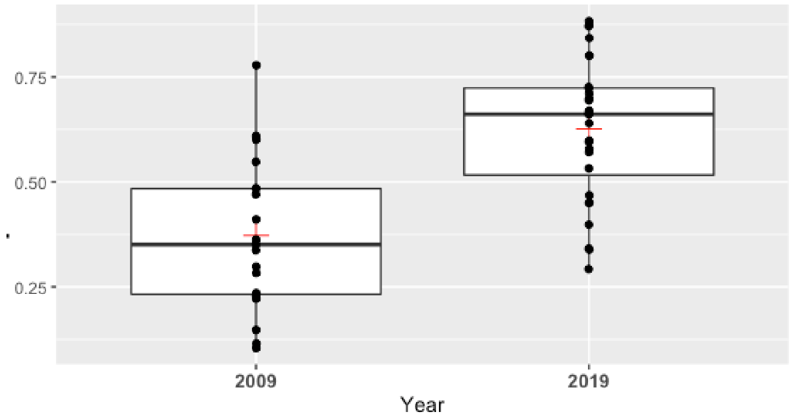


Figure 1: Proportion of prose devoted to statistical reporting per article in 2009 and 2019.

We also noted a statistically significant increase in the number of distinct statistical techniques used (*p* = 0.0075, *d* = -0.83), as shown in Figure 2; outliers are marked by a red “x”.

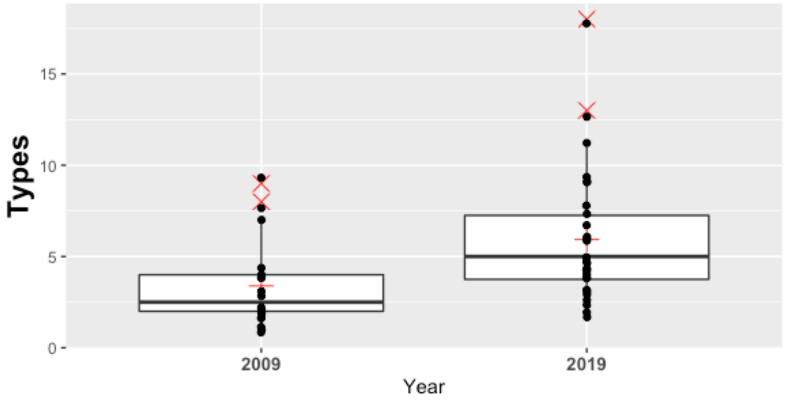


Figure 2: Number of distinct statistical techniques per article in 2009 and 2019.

The mean number of distinct statistical techniques used has risen from 3.5 (*SD* = 2.3) in 2009 to 5.9 (*SD* = 3.7) in 2019, with three studies using more than 10 distinct techniques in 2019. A closer look at the techniques used showed that there was a certain degree of overlap between the years, primarily for descriptive techniques (e.g., means, SD), measures of association strength (e.g., *log-likelihood, MI*), and tests of frequency differences (e.g., *Chi-square tests*). Techniques unique to 2009 included, for example, the *Pearson residual*, whereas studies in 2019 used techniques such as *Hierarchical Cluster*

*Analysis*, *Linear mixed models* and *Classification and Regression Trees (CART)*.

On the one hand, these changes could be seen as a step in the right direction in that the field is showing signs of increased statistical literacy and better familiarity with advanced techniques. On the other hand, we have evidence to suggest that there is a second, related trend in play: the increased focus on statistics appears at least to some degree to have come *at the expense* of linguistic description and analysis.

In theory, there is of course no inherent opposition between statistical and linguistic analysis; statistical methods are used to test linguistic data and help the analyst draw conclusions about language. Nonetheless, we noted a possible de facto opposition: the proportion of prose devoted to statistical reporting was found to be negatively correlated with number of text excerpts and linguistic examples included ($r$ = -0.56). To a varying degree, articles that have a strong statistical focus appear to abstract away from language without explicitly returning to it. This means that the increase in proportion of words devoted to statistical reporting does not seem to be attributable merely to an augmented need for more words to describe increasingly advanced statistical techniques.

All in all, while we do not question the usefulness of statistical methods for corpus linguistics research, we find it troubling that linguistic analysis seems to be increasingly backgrounded. In this talk and in the forthcoming article (Larsson, Egbert, & Biber, forthcoming), we propose that corpus linguists can combine statistical analysis and linguistic analysis in a more balanced way.

## *References*

Gass, S., Loewen, S., & Plonsky, L. (2020). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, 1–14.

Larsson, T., Egbert, J., & Biber, D. (forthcoming). On the status of statistical reporting versus linguistic description: A ten-year perspective. *Corpora*, *17*(1).

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research, 3(1),* 61–94.

**Laube, Alexander**

***University of Regensburg***

**alexander.laube@ur.de**

Type of Contribution: Full Paper

# A corpus-based approach to style

## Stylistic variation in Bahamian English

*Abstract*

Ever since Labov's (1966) famous New York study, style has been a central concept in sociolinguistic theory. While the concept was originally restricted to inter-speaker variation in the Labovian sense, i.e., variable language use depending on, for example, the formality of the particular speech situation, researchers now put an emphasis on speakers not simply being "passive and stable carriers of dialect, but […] stylistic agents, tailoring linguistic styles in ongoing and lifelong projects of self-construction and differentiation" (Eckert, 2012, pp. 97-98). In other words, in addition to accommodating the nature of the particular speech situation, speakers create specific images of self when talking to each other and this results in observable stylistic variation. Unfortunately, however, corpus linguistic methods are often unsuited to examine phenomena such as style-shifts on the level of individual speaker, especially when dealing with conversational data. Thus, the present study, looking into stylistic variation in the speech of Bahamians, explores in how far inter- as well as intra-speaker variation can be measured quantitatively, putting two methodological approaches to the test.

### Data and method

The study builds on a corpus of conversational data from The Bahamas that represents the intermediate section of the creole continuum[1]. It comprises

[1] As is the case in most creole-speaking societies, the linguistic situation in The Bahamas is one that "can be best described as a continuum of overlapping varieties of English, ranging from a creole […] (the basilect) to a variety of English whose grammatical differences from the

mesolectal creole data from the late 1990s (cf. Hackert, 2004) as well as a range of conversations from the Bahamas subcomponent of the *International Corpus of English*, i.e., personal conversations and broadcast discussions/interviews, which (to varying degrees) represent the acrolectal end of the continuum. The corpus totals at around 360,000 words[2]. As the paper is mainly concerned with method, the two methodological approaches will now be dealt with in turns.

## The variationist approach

The study of stylistic variation has a long tradition in variationist sociolinguistics (VSLX). Initially restricted to attention paid to speech (Labov, 1966), style has over the years grown into one of the field's most important variables, receiving much attention in sociolinguistic theory (cf. Eckert, 2012). In this paper, building on Kiesling's (2009, p. 172) idea "that people's primary way of organizing interaction […] is through stances" and that stance is thus at the heart of stylistic variation, I attempt to model the effects of style on variable copula use in the speech of Bahamians, i.e., copula deletion and present and past *be* leveling, as in *I's a teacher* or *We was running*.

As is standard practice in VSLX, the tokens belonging to the respective envelopes of variation were extracted and manually coded for a variety of linguistic, social and, of course, stylistic factors. The linguistic factors are mainly concerned with surrounding grammatical and/or phonetic environment; social factors are restricted to text type and individual speaker. As regards style, I again follow Kiesling (2009) and Holmes-Elliot & Levon (2017) in assuming that stance is directly connected to linguistic style and can be coded for indirectly using a range of speech activities, thus providing "a clear method for operationalizing stance in a replicable and objective fashion" (Holmes-Elliot & Levon, 2017, p. 1054). For my analysis of copula deletion, I applied a slightly modified version of their coding scheme (cf. Holmes-Elliot & Levon, 2017, p. 1056) to my data and the dataset was then further processed and analyzed in

standard English spoken elsewhere are negligible (the acrolect)" (Holm & Shilling, 1982, p. ix).

[2] Note that not all of the analyses described in this paper are currently based on the entire dataset.

*R*.[3] A generalized mixed effects model using the *glmer* function from the *lme4* package was fitted and the final model contained three simple main effects (preceding environment, following environment, negation) as well as a two-way interaction of text type and stance as fixed effects; individual speaker was added as a random intercept.

### Multiple Correspondence Analysis

For the second approach, I subjected my data to a recent offshoot of the classic *Multidimensional Analysis* (MDA) (Biber, 1988), i.e., short-text MDA (cf. Clarke & Grieve, 2019). This new methodological approach records the presence or absence of a range of grammatical features in each text, resulting in a categorical data matrix which is then subjected to a *Multiple Correspondence Analysis* (MCA). The approach is thus targeted specifically at analyzing short texts and it seems reasonable to assume that it could be fruitfully applied not only to analyzing shorter texts, but, for example, to look into common patterns of variation at the individual utterance level, thus zooming in on intra-textual or intra-speaker variation.

In order to conduct the analysis, I had to subject my corpus to a range of modifications. First, the corpus texts were semi-automatically converted into a "turn-by-turn" format, i.e., text units comprising one speaker turn each. Next, the data were POS-tagged using the *Multidimensional Analysis Tagger* (Nini, 2015) and post-processed to, among other things, incorporate variety-specific grammatical features like preverbal TMA markers, negative *ain't* and lack of past marking (mainly using regular expressions and *SarAnt* (Anthony, 2016)). The dataset was then further processed and analyzed in *R*; the MCA was conducted, for the most part following the supplementary markdown provided by Clarke & Grieve (2019). For the analysis, I reduced my dataset, excluding all text units, i.e., utterances, with less than five words. The interquartile range of the remaining data was 8 to 27 words per text unit. In addition, all irrelevant POS-tags as well as tags which occurred in less than 2% of all text units were

[3] In addition to the R software (R Core Team, 2020), a number of packages were used for data preparation, analysis and plotting: *car* (Fox & Weisberg, 2019), *dplyr* (Wickham et al., 2020), *effects* (Fox, 2003), *FactoMineR* (Le et al,. 2008), *ggplot2* (Wickham, 2016), *lme4* (Bates et al., 2015), *phia* (Rosario-Martinez, 2015), *sjPlot* (Lüdecke, 2020).

excluded, leaving me with 35 tags. Both word count and text type (or register) were included as supplementary continuous or categorical variables in the analysis.

## Results

First of all, the study finds that, in addition to grammar and register, the occurrence of non-standard variants is constrained by distinct speech activities, which correspond to linguistic styles or stances taken by the speaker. As illustrated by Figure 1 below, social speech situations like joking or gossiping clearly favor the application of non-standard variants like ZERO *be*, whereas informational activities or discourse management do not show this effect. Most strikingly, this pattern can be traced through all three text types and the effect of the social stance dimension appears to take precedence over text type. Nevertheless, the corresponding confidence intervals also show that the interviews, i.e., the creole sample, display noticeable variation in the stance variable; there is, however, strong evidence for social constraints in this text category, in that the distribution of ZERO *be* across the individual speakers mirrors previous findings (cf. Hackert, 2004).

As regards the short-text MDA, the findings of the present study really raise more questions than they provide answers. I conducted the analysis following Clarke & Grieve (2019) and closely examined the first four dimensions returned by the MCA, which together account for 87.7% of the variance in the data[4].

[4] Like Clarke & Grieve (2019, p. 8), I calculated the variance "using the standard adjustment for MCA".

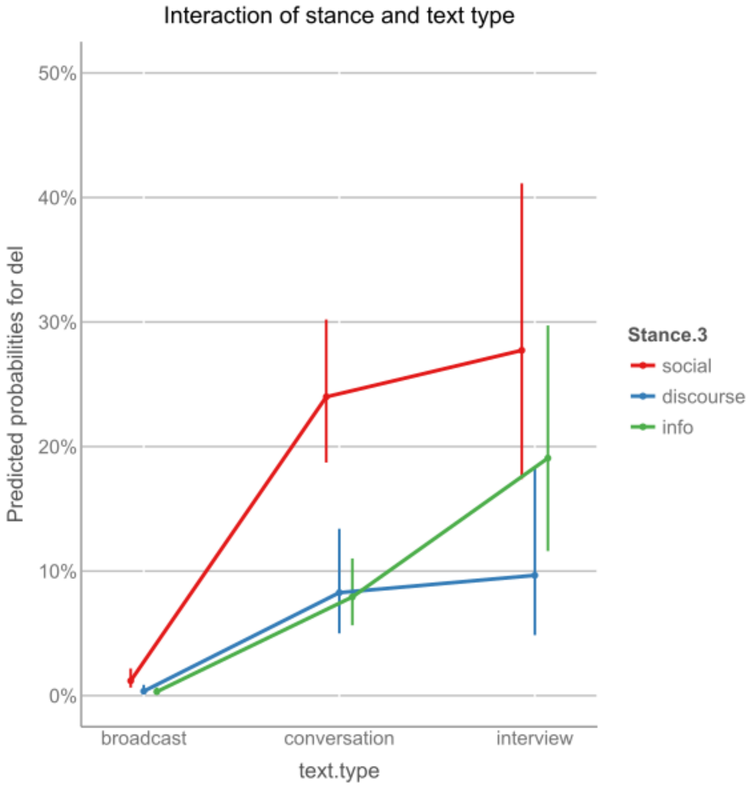


Figure 1: Effects plot for the interaction of stance and text type.

Just like in their study, Dimension 1 almost exclusively represents text length, i.e., the longer the text the more likely it is that one of the linguistic features would occur, and the other dimensions cover some of the stylistic variation. However, only around 4% of this variance fall on dimensions two to four, meaning that the stylistic constraints are not very strong for the Bahamian conversational data. In any case, this needs further exploration.

## *References*

Anthony, L. (2016). *SarAnt*. Version 1.1.0. Tokyo: Waseda University. Retrieved from: https://www.laurenceanthony.net/software.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Clarke, I., & Grieve, J. (2019). Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLoS ONE, 14*(9), e0222062.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology, 41*, 87-100.

Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression*. Thousand Oaks, CA: Sage.

Hackert, S. (2004). *Urban Bahamian creole: System and variation*. Amsterdam: Benjamins.

Holm, J. A., & Shilling, A. W. (1982). *Dictionary of Bahamian English*. Cold Spring, NY: Lexik House.

Kiesling, S. F. (2009). Style as stance: Stance as the explanation for patterns of sociolinguistic variation. In A. M. Jaffe (Ed.), *Stance: Sociolinguistic perspectives* (pp. 171-194). Oxford: Oxford University Press.

Labov, W. (1966). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software, 25*(1), 1-18.

Lüdecke, D. (2020). *sjPlot: Data visualization for statistics in social science*. R package version 2.8.3.

Nini, A. (2015). *Multidimensional analysis tagger*. Version 1.3. Retrieved from: http://sites.google.com/site/multidimensionaltagger.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: https://www.R-project.org/.

Rosario-Martinez, H. de. (2015). *phia: Post-Hoc Interaction Analysis*. R package version 0.2-1. Retrieved from: https://CRAN.R-project.org/package=phia.

Wickham, H. (2016). *ggplot2. Elegant graphics for data analysis*. New York: Springer.

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A grammar of data manipulation*. R package version 0.8.5. Retrieved from: https://CRAN.R-project.org/package=dplyr.

**Le Foll, Elen**

***Osnabrück University***

**elefoll@uos.de**

Type of Contribution: Full Paper

## Exploring the registers of school EFL textbooks

Using multi-dimensional analysis

*Abstract*

This study sets out to uncover the specific combinations of linguistic features that characterize "the kind of synthetic English" (Römer, 2004, p. 185) that pupils are exposed to via their school textbooks using Multi-Dimensional Analysis (MDA; Biber, 1988; Conrad & Biber, 2001/2013; Sardinha & Biber, 2014; Sardinha & Pinto, 2019). To this end, a Textbook English Corpus made up of 43 English as a Foreign Language (EFL) coursebooks used in lower secondary schools in France, Germany and Spain is mobilized. All the texts of the textbooks were manually annotated for the following intra-textbook registers: Conversation (including the transcripts of the accompanying audio and video materials), Personal Communication (letters, diary entries…), Informative texts, Song & Poetry, Fiction, and Instructional texts.

First, an additive MDA (Sardinha et al., 2019) was carried out to map out these six Textbook English registers against Biber's (1988) model as a "base-rate knowledge of English register" (Nini, 2019, p. 70). An adapted version of the MAT tagger (Nini, 2019) was used to tag and count all linguistic features. To boost comparability, three additional target learner language corpora were also tagged, analyzed and mapped onto Biber's (1998) dimensions following the same procedure as for the textbook register subcorpora. These are:

(1) The Spoken BNC2014 (1,251 conversations from the UK; Love et al., 2017)

(2) A custom-built Youth Fiction Corpus (1,191 samples from 300 novels for teenagers and young adults)

(3) A custom-built Informative Texts for Teens Corpus (1,414 individual texts from 22 web domains including *dogonews.com*, *factmonster.com* and *whyfiles.org*)

As expected, among the textbook registers, textbook conversation displays the highest scores on Biber's (1988) first, Involved vs. Informational Production, dimension. Nonetheless, when compared to the Spoken BNC2014 scores on this dimension, school EFL textbook dialogues clearly underrepresent many of the features typical of naturally occurring spontaneous conversation, such as private verbs (e.g. *know*, *mean*, *think*), contractions, WH-clauses, demonstrative pronouns and hedges. Informative texts from textbooks also differ quite substantially from those in the Informative Texts for Teens Corpus but, unlike textbook dialogues, the informative texts of the more advanced textbooks, in which more complex structures are introduced to the learners, are less strikingly different. The mean Dimension 1 scores of textbook fiction and the Youth Fiction texts are not significantly different. This suggests that their proportions of narration to fictional speech are similar (cf. Egbert & Mahlberg, 2020). Textbook fiction is similar to the Youth Fiction Corpus on all dimensions and, indeed, many of the fiction texts featured in school EFL textbooks are extracts of the kind of novels featured in the Youth Fiction corpus. On Biber's (1988) third dimension, textbook instructions and explanations are found to be particularly "explicit" as opposed to "situation-dependent". This is partly due to the extensive use of phrasal coordination in formulaic trigrams such as *ask and answer*, *listen and check*, *words and phrases, true and false*, etc.

Second, a full MDA was carried out on the textbook and target learner language corpora. To this end, 84 grammatical, lexical and semantic features were tagged and counted across the 5,805 texts, 72 of which were later entered in the exploratory factor analysis (KMO = 0.91). Four factors were extracted, accounting for 41% of the total variance. The first factor corresponds to the oral-literate dimension that has emerged as an almost universal dimension across many different registers and languages (Biber, 2014). This first, 'Involved vs. Informational' dimension accounts for 21% of the variance. Among the textbook registers, Conversation and Personal Communication score highest. The large gap between the mean score of the textbook dialogues ($\bar{x}$ = 6.73, *std* = 7.88) and that of the Spoken BNC2014 ($\bar{x}$ = 33.6, *std* = 5.58) points to major linguistic differences between these two types of “spoken” English. This is largely due to the more nominal style of the textbook dialogues. They also feature far fewer pragmatic markers, filled pauses, interjections, negated verbs, question tags and predictive modals, among others. The second dimension to emerge is interpreted as ‘Complex Edited vs. Direct Personal Exchange’. Informative texts score highest on this dimension. Unsurprisingly, textbook informative texts are found to be less complex than those targeted at L1 English-speaking teenagers. The third dimension is very similar to Biber’s (1988) ‘Narrative’ dimension and also includes past tense, third person pronouns, as well as phrasal, activity and aspect verbs as features contributing to high scores. The fiction and conversation reference corpora are found to be more narrative than their corresponding textbook subcorpora, whilst textbook informative language is more narrative than that of the informative websites for teenagers. Finally, the fourth dimension clearly exposes the textbooks’ instructions and explanations as a very distinct register, which scores much higher on this ‘Instructional’ dimension than any other textbook or target learner language register. This factor only includes features with positive loadings: verbs in the imperative, mental verbs (e.g. *think, know, read, choose*), second person pronouns, WH pronouns, WH clauses, communication verbs (e.g. *answer, say, write*) and WH determiners.

In conclusion, whilst Textbook English does present some register variation, both MDAs show that register-based variation is far more restricted than in Biber’s (1988) general English corpus or in the target learner language data examined as part of this study. Textbook conversation is found to be a rather poor representation of naturally occurring conversation. By contrast, textbook fiction shares many of the characteristics of novels targeted at teenage L1 readers. This study highlights the potential of MDA to pinpoint the defining linguistic features that characterize Textbook English registers as compared to naturally occurring registers with similar communicative purposes and target audience.

## *References*

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in contrast*, *14*(1), 7-34.

Conrad, S., & Biber, D. (Eds.). (2013). *Variation in English: Multi-dimensional studies*. Routledge. (Original work published 2001).

Egbert, J., & Mahlberg, M. (2020). Fiction - one register or two? Speech and narration in novels. *Register Studies*, *2*(1), 72-101.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics*, *22*(3), 319-344.

Nini, A. (2019). The multi-dimensional analysis tagger. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 67-96). London: Bloomsbury.

Römer, U. (2004). A corpus-driven approach to modal auxiliaries and their didactics. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 185-199). Amsterdam: John Benjamins.

Sardinha, T. B., & Biber, D. (Eds.). (2014). *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*. Amsterdam: John Benjamins.

Sardinha, T. B., & Pinto, M. V. (2019). *Multi-dimensional analysis: Research methods and current issues*. London: Bloomsbury Academic.

Sardinha, T. B., Pinto, M. V., Mayer, C., Zuppardi, M. C., & Kauffmann, C. H. (2019). Adding registers to a previous multi-dimensional analysis. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 165-188). London: Bloomsbury.

**Leclercq, Benoît**

***Université de Lille***

**benoit.leclercq@univ-lille.fr**

Type of Contribution: Full Paper

## From modals to modal patterns

An n-gram analysis of can, could and be able to

*Abstract*

The goal of this paper is to present the results of a corpus analysis aimed at identifying a number of modal patterns with *can*, *could* and *be able to*. Both onomasiological and semasiological approaches have been used to look at the three modal expressions, either viewing them as a paradigm for the expression of 'ability' or focusing on their individual idiosyncratic properties (cf. Coates, 1983; Palmer, 1990; Westney, 1995; Facchinetti, 2000, 2002; Aijmer, 2004; Collins, 2009; *inter alia*). In either case, the discussion remains typically focused on the verbs alone, however. Even though the context in which these verbs are used impact on the meaning communicated, a systematic analysis of the lexical cooccurence patterns, and their semantic and pragmatic potential, is lacking. Yet, as Firth points out, "you shall know a word by the company it keeps!" (in Palmer, 1968, p. 179) The main goal of this paper is to present the results of a new corpus analysis which precisely aimed at identifying such patterns. It will be shown that a qualitative analysis of these patterns is directly relevant to the study of the three modal verbs.

Using Mark Davies' (2008) interface, the methodology developed by Cappelle and Depraetere (2016) was applied to the Corpus of Contemporary American English (COCA). I looked for direct collocates to the left and right of each modal and extracted all bigrams that occurred with a frequency of at least 250 and a *mutual information* (MI) score of minimum 3. I obtained sequences such as *can go*, *could use* and *longer able*. For each of the bigrams found, I looked again for direct collocates to their right and left (with a minimum

frequency of 250 and MI-score of 3), and found three-word sequences like *can go wrong*, *could use some* and *no longer able*. These patterns then served as search items for the retrieval of yet larger patterns, and the procedure was then repeated again and again until no further patterns could be found. This recursive procedure resulted in the extraction of a total of 1,640 modal patterns. A selection of these patterns will be discussed here, paying close attention to their semantic and pragmatic features.

Compared to *can* and *could*, the periphrastic form *be able to* was found in a small number of n-grams (287), most of which illustrate the wide variety of syntactic contexts in which it is the only possible alternative (e.g. *being able to*, *had been able to*, *'ll never be able to*, *might not be able to*, *we should be able to*). Few conclusions about the semantics and pragmatics of the modal expression can thus be drawn. Two sets of n-grams did reveal interesting properties though: uses of *be able to* in a perfect tense (SUBJ HAVE *been able to* VP) and in the past tense (SUBJ BEpast *able to* VP). A qualitative analysis shows that these patterns are typically used to express the modal meaning of 'opportunity' (cf. Depraetere & Reed, 2011) and also entail actualization (Bhatt, 1999):

(1) The argument is that Putin has won in Syria, that he **has been able to** get a foothold in the Middle East because of U.S. policy. (COCA, spoken)

(2) I **was able to** get into the house with surprising ease, Sam. The library window was left unlocked. (COCA, written)

This type of examples was to be expected since ‘actualized opportunity’ has previously been shown to be a conventional feature of *be able to* (Leclercq & Depraetere, forthcoming.). Nevertheless, given the generalization processes required to storing constructions, I will show that the data presented here also raises some questions about the level at which 'actualized opportunity' may be represented.

The modals *can* and *could* were found in a much bigger number of n-grams (844 and 509 respectively). A qualitative analysis reveals that many of these n-grams show idiosyncratic properties and therefore obtain their own *construction* status (Goldberg, 2006, p. 5). Consider the sequences in bold in the following examples:

(3) Asked whether he had done such a thing, Walley said “Lord, no. I got better sense than that.” His father, Roy, ***could* not be reached for comment**. (COCA, written)

(4) Sweet loyal Jack. **I *can*’t tell you how** good it is to see you. (COCA, written)

(5) We don’t fight these wars to win them anymore. **I don’t think we *can*** risk American treasure and lose that treasure going forward any further. (COCA, spoken)

In example (3), *could* communicates ‘opportunity’ and the sequence in which it occurs typically implies that the subject referent is unwilling (rather than unable) to comment. In (4), the pattern with *can* is used to index the speaker's (in)ability to be specific about quantities that are strickingly high, and it is used to emphasize one's positive (or negative) appreciation of the situation. In example (5), the modal *can* is used in a context (negative raising with *think*) that weakens its modal value, and the speaker communicates a suggestion not to perform the action denoted by the embedded VP. What these examples are therefore meant to show is that the interpretation of sentences that contain *can* or *could* does not only depend on the semantics of the modal verbs alone and/or the extra-linguistic context in which they are found: the meaning of the modal expressions is also greatly determined by the larger sequences (i.e. modal idioms) in which they are stored.

On the basis of this observation, the concluding part will close the discussion by arguing that greater descriptive accuracy can be achieved when modals are therefore construed as part of complex networks of inter-related constructions rather than in terms of isolated nodes. It will be shown, for instance, that *can* is used in sequences (e.g. (5)) that otherwise seem typical of necessity modals (e.g. (6) to (9)).

(6) It was kind of Anne to think of it and not condemn us, but - well we’ve managed so far being discreet about it, and **I don’t think we *should*** change things now. (Cappelle, Depraetere & Lesuisse, 2019:232)

(7) Well, I agree with you. **I don’t think we *have to*** give up any liberties. But we’ve got to work smarter and not just harder. (COCA, spoken)

(8) And let me be clear on this. **I don't think we *need to*** have our boots on the ground. The Kurds are a very viable military force. (COCA, spoken)

(9) **I don't think we *ought to*** attack the Social Security system. It is the last line of defense that Americans have when they lose their pensions. (COCA, spoken)

The semantic proximity of *can* with the necessity modals in those examples is best understood when taking into account the construction in which the modal is used rather than by looking at *can* in isolation. This shows that the lexico-grammatical environment of a modal verb can thus give valuable insights into its semantic and pragmatic properties.

## *References*

Aijmer, K. (2004). The semantic path from modality to aspect: *Be able to* in a cross-linguistic perspective. In H. Lindquist & C. Mair (Eds.), *Corpus approaches to grammaticalization in English* (pp. 57-79). Amsterdam: John Benjamins.

Bhatt, R. (1999). *Covert modality in non-finite contexts*. PhD thesis. University of Pennsylvania.

Cappelle, B., & Depraetere, I. (2016). Response to Hilpert. *Constructions and Frames, 8*(1), 86-96.

Cappelle, B., Depraetere, I., & Lesuisse, M. (2019). The necessity modals *have to*, *must*, *need to* and *should*: Using n-grams to help identify common and distinct semantic and pragmatic aspects. *Constructions and Frames, 11*(2), 220-243.

Collins, P. (2009). *Modals and quasi-modals in English*. Amsterdam/New York: Rodopi.

Coates, J. (1983). *The semantics of the modal auxiliaries*. London/Canberra: Croom Helm.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA): 600 million words, 1990-present*. Retrieved from: https://www.english-corpora.org/coca/.

Depraetere, I., & Reed, S. (2011). Towards a more explicit taxonomy of root possibility. *English Language and Linguistics, 15*(1), 1-29.

Facchinetti, R. (2000). *Be able to* in present-day British English. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory* (pp. 117-130). Amsterdam: Rodopi.

Facchinetti, R. (2002). *Can* and *could* in contemporary British English: A study of the ICE-GB Corpus. In P. Peters, P. Collins & A. Smith (Eds.), *New frontiers of corpus research* (pp. 229-246). Amsterdam: Rodopi.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Leclercq, B., & Depraetere, I. (forthcoming). Making meaning with *be able to*: modality and actualisation. *English Language and Linguistics*.

Palmer, F. (1968). *Selected papers of J.R. Firth 1952-59*. London: Longman.

Palmer, F. (1990). *Modality and the English modals*. 2nd edition. London: Longman.

Westney, P. (1995). *Modals and periphrastics in English*. Tubingen: Niemeyer.

**Lehmann, Claudia**

***University of Bremen***

**claleh@uni-bremen.de**

Type of Contribution: Full Paper

# Using multimodal corpora to study the ironic tone of voice

## Opportunities and challenges

*Abstract*

Research on the ironic tone of voice in natural contexts has ever been challenging. Given that available speech corpora are not tagged for irony, researchers have often resorted to either experimental research designs (Cheang & Pell, 2008; Rockwell, 2000) or smaller, sampled corpora (Attardo et al., 2003; Bryant, 2010). These studies show that there is a considerable difference between the prosodic patterns of elicited and non-elicited irony. Elicited irony is reported to be marked by a longer duration (Cheang & Pell, 2008; Mauchand et al., 2018; Rockwell, 2000), a lower mean pitch (Cheang & Pell, 2008; Rockwell, 2000), reduced pitch variability (Cheang & Pell, 2008; Mauchand et al., 2018), a higher mean intensity (Rockwell, 2000), and a greater amount of noise (Cheang & Pell, 2008). Non-elicited irony is also reported to be marked by a longer duration (Bryant, 2010; Rockwell, 2007), but, in contrast to elicited irony, marked by higher mean pitch (Bryant & Fox Tree, 2005; Rockwell, 2007), more pitch variability (Rockwell, 2007) and more intensity variability (Bryant & Fox Tree, 2005).

As becomes apparent from this overview, research on the ironic tone of voice remains inconclusive. The most striking puzzle is the considerable divergence between studies using elicited irony in laboratory settings and those analyzing irony in natural settings. Bryant (2010) proposes that the actors who produced the elicited forms of irony may have used a stylized prosodic pattern while naturally occurring irony may be marked in several, related ways. Given that studies analyzing naturally occurring irony have been working with rather

small sample sizes, which is due to the fact that they heavily rely on novel, ad hoc ironies, which are difficult to find in large, general corpora, the existence of several tones of voices for irony is only hypothetical. There are, however, quite a few examples of what Giora (2003) calls ‘salient’ ironies, which have received growing attention recently. These formally more or less fixed constructions allow analyzing the prosodic pattern(s) associated with irony on a larger scale and the research report that follows provides such an analysis.

For the study reported here, the syntactic string *Tell me about it,* which allows for an ironic and a nonironic reading, was searched for in the multimodal *NewsScape Library of Television News Broadcasts* (Steen & Turner, 2013) from 22/11/2016 to 13/03/2020. This search resulted in 1019 hits in total, including many duplicates, syntactically integrated occurrences, and results that are transcribed incorrectly, all of which have been removed from further analyses. Results including considerable overlaps of the target utterance with utterances by other speakers have been discarded, too. The remaining 269 observations were annotated for interpretation (ironic, nonironic), data type (scripted, nonscripted), speaker gender (male, female), and a number of acoustic variables (duration, mean pitch, pitch range (max-min), pitch sd, harmonics-to-noise ratio, mean intensity, intensity range (max-min)). The acoustic variables were measured with *Praat* (Boersma & Weenink 2019) using the autocorrelation method. Due to the low frequency of scripted, nonironic hits (N=4), scripted data was excluded from further analyses, leaving 226 remaining observation.

A series of two-way ANOVAs was performed using gender and interpretation as factors and the acoustic measures as dependent variables. The results of these show that there is a main effect of gender on mean pitch, pitch range, pitch sd, and harmonics-to-noise ratio; all of which quite predictable results. More interesting, though, is the finding that there were main effects for gender and interpretation on duration ($p < 0.001$) with small to medium effect sizes (partial $\eta^2 = 0.062$ for gender and 0.096 for interpretation), but no interaction. Furthermore, there was an interaction effect on mean intensity ($p < 0.01$, $\eta^2 = 0.045$), but no main effects.

The dataset was then explored using the unsupervized k-means cluster analysis algorithm (Wong & Hartigan, 1979). The gap statistic method

(Tibshirani et al., 2001) suggested the optimal number of clusters be four and so k-means clustering was computed with four clusters using duration, mean pitch, pitch sd, and mean intensity. The cluster means are summarized in Table 1:

| | **t** in ms | mean **F0** in Hz | **F0 sd** in Hz | mean **I** in dB |
|---|---|---|---|---|
| 1 | 657 | 253 | 80.6 | 63.5 |
| 2 | 786 | 172 | 38.1 | 64.0 |
| 3 | 570 | 151 | 28.5 | 58.4 |
| 4 | 552 | 148 | 29.9 | 66.8 |

Table 1: Means of the four clusters.

The resulting cluster vector was then treated as factor variable and was tested for associations with gender and interpretation, using χ² as an association measure. The results are visualized in Figure 1 and 2:

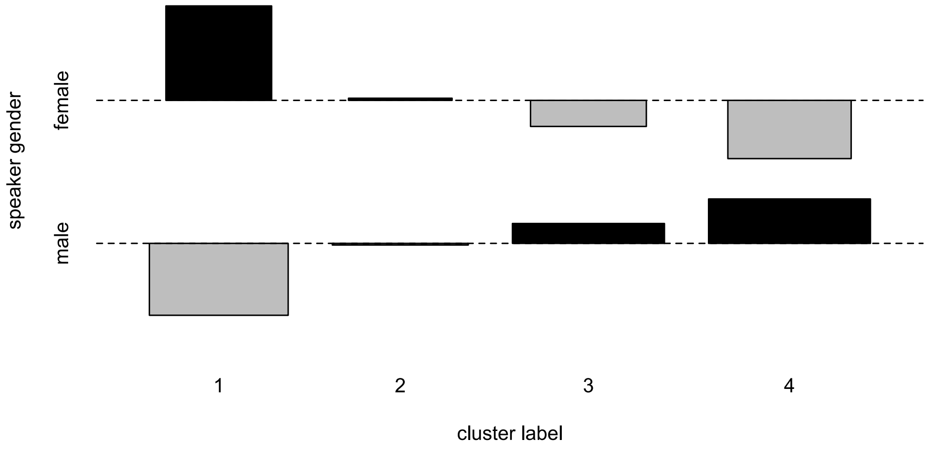


Figure 1: Association plot of CLUSTER ~ GENDER.

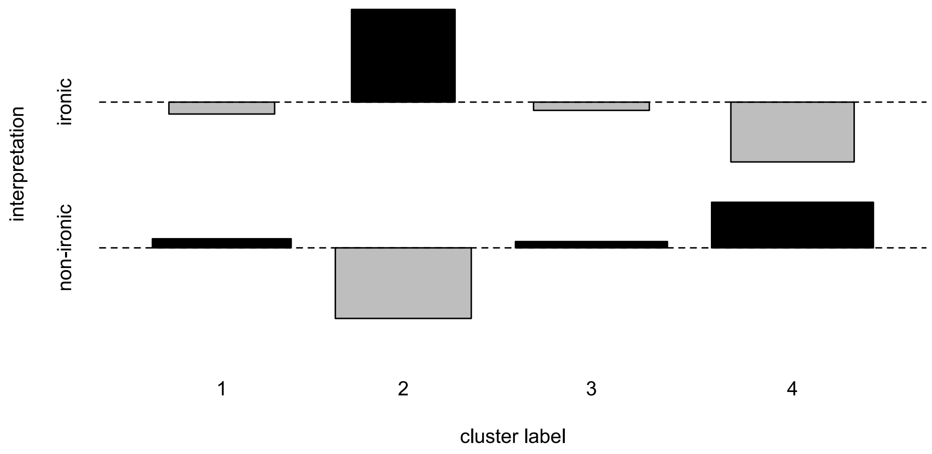


Figure 2: Association plot of CLUSTER ~ INTERPRETATION.

These results suggest the following:

- Cluster 1 being associated with women, but no particular interpretation.
- Cluster 2 being associated with irony, but no particular gender.
- Cluster 3 being weakly associated with men, but no particular interpretation.
- Cluster 4 being associated with men and a nonironic interpretation.

The discussion section of the paper is divided into three parts. In the first part, conclusions on the ironic tone of voice will be drawn. It will be argued that the existence of several tones of voices is rather unlikely and that the divergent findings of previous studies might be due to gender being a confounding factor, which has not been controlled for in sufficient detail. The second part of the discussion will highlight methodological challenges, in particular sound quality issues and intensity measures. And, finally, in the third part, the opportunities the *NewsScape Library of Television News Broadcasts* offers are summarized, including its large database and its potential of providing insights into meaning-making processes.

## *References*

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor - International Journal of Humor Research, 16*(2), 243-260.

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer. Amsterdam.

Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice? *Language and speech, 48*(3), 257-277.

Bryant, G. A. (2010). Prosodic contrasts in ironic speech. *Discourse Processes, 47*(7), 545-566.

Cheang, H. S., & Pell, D. M. (2008). The sound of sarcasm. *Speech Communication, 50*, 366-381.

Giora, R. (2003). *On our mind: Salience, context, and figurative language*. Oxford: Oxford University Press.

Mauchand, M., Vergis, N., & Pell, M. (2018). Ironic tones of voices. *Paper presented to the 9th International Conference on Speech Prosody 2018*.

Rockwell, P. (2000). Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic Research, 29,* 483-495.

Rockwell, P. (2007). Vocal features of conversational sarcasm: A comparison of methods. *Journal of Psycholinguistic Research, 36*, 361-369.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*, 411-423.

Wong, M. A., & Hartigan, J. A. (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society, 28*(1), 100-108.

**Li, Yingyu**

***Xi'an Jiaotong University***

**hiwendy@xjtu.edu.cn**

**Zhang, Jian**

***Xi'an Jiaotong University***

**zhangjane@stu.xjtu.edu.cn**

Type of Contribution: Work-in-Progress Report

# The change of emotion in translations from Chinese to English

A corpus-based comparison between L1 and L2 translation

*Abstract*

There are scholars and organizations holding the idea that translators should translate into their mother tongue, i.e., L1 translation (Graham, 1965; Kelly, 1979; Newmark, 1988; Baker, 2009), however, many people start to reconsider the value of translating into the "other direction", or L2 translation (Snell & Crampton, 1989; Campbell, 1998; Pokorn, 2005; Huang, 2011; Wang, 2015), even when "there is no consensus about the terminology used to refer to directions in translation" (Beeby, in Baker, 2009).

Emotional words are important in literary works for their power to express subtle feelings, which causes challenges for not only understanding but expressing as well. Therefore, it would be an ideal window of observing the differences between L1 and L2 translation, through which the three questions will be explored in this research:

(1) What are the emotional words translated in L1 and L2 versions?

(2) In which aspects are they similar or different?

(3) How close, or far, are the translations?

We have a ready Chinese-English translational corpus called PCSL. However, effective comparison can only be made when the originals are the same (Huang, 2011), so works with both L1 and L2 translations are considered as ideal, and in order to keep the corpus balanced and representative, short stories can be kept full while novels should be cut short. Thus, a sub-corpus was created, which included 68,894 Chinese characters and 105,650 English words, which is composed of 3,739 pairs of sentences. The detailed information is shown in Table 1:

| Original | Tokens (Characters) | Translation | Translator | Tokens (words) |
|---|---|---|---|---|
| *Chuiniu* | 6,778 | *Talking Bull* | Lixia Du | 5,421 |
| | | *Shooting the Bull* | Zachary Haluza | 4,927 |
| *Li Shisan Tuimo* | 10,038 | *A Tale of Li Shisan and the Millstone* | Jianchong Nan | 6,915 |
| | | *Li Shisan Worked the Millstone* | Philip Hand | 7,260 |
| *Qinqiang (excerpt)* | 12,412 | *Shaanxi Opera* | Zongfeng Hu | 9,976 |
| | | *Shaanxi Opera* | Dylan King | 10,421 |
| *Feidu (excerpt)* | 18,113 | *The Abandoned Capital* | Zongfeng Hu | 13,957 |
| | | *Ruined Capital* | Howard Goldblatt | 12,625 |
| *Gaoxing (excerpt)* | 21,553 | *Happy* | Zongfeng Hu | 18,704 |
| | | *Happy Dreams* | Nicky Harman | 15,444 |

Table 1: General information of PCSL sub-corpus.

As it is known, the Chinese language is very powerful in creating new words with the limited characters, so by referring to the HowNet Wordlist (Sentiment), we selected four characters, namely, ***ai*4 (love), *qin*1 (intimacy), *ku*3 (bitterness) and *hen*4 (hatred)**, two of which are generally accepted as having more positive emotions while the other two as having more negative emotions. When the four characters are used as search words, all the related emotional words in the original texts can be found. In this way, we have got 186 pairs of sentences, 93 from L1 and L2 translations respectively, with a total of 41 different emotional words there. The high frequency words are shown in table 2.

| L1 Translation | Frequency | L2 Translation | Frequency |
|---|---|---|---|
| **love** | 7 | love | 8 |
| **hate** | 5 | hate | 6 |
| **like** | 5 | bitter | 3 |
| **kiss** | 3 | fond | 3 |
| **care** | 2 | kiss | 3 |
| **dear** | 2 | like | 3 |
| **relative** | 2 | relative | 3 |
| **want** | 2 | beloved | 2 |
| | | care | 2 |
| | | complain | 2 |
| | | dear | 2 |
| | | hard | 2 |
| | | odd | 2 |
| | | poor | 2 |
| | | resent | 2 |

Table 2: High frequency words in L1 and L2 translations.

With reference to the *Modern Chinese Dictionary*, *Oxford Advanced English Chinese Dictionary*, and the corpus of COCA, the 93 emotional words and their 186 equivalents from L1 and L2 translations were tagged as “positive”, “negative”, or “neutral”. The results are as follows (Table 3):

| | **Positive** | | | **Negative** | | | **Neutral** | | | **Not Translated** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Originals | L1 | L2 | Originals | L1 | L2 | Originals | L1 | L2 | L1 | L2 |
| **love** | 25 | 20 | 22 | 1 | 2 | 2 | 2 | 4 | 2 | 2 | 2 |
| **intimacy** | 15 | 11 | 12 | 0 | 0 | 0 | 12 | 11 | 10 | 5 | 5 |
| **bitterness** | 2 | 2 | 1 | 16 | 8 | 15 | 0 | 4 | 1 | 4 | 1 |
| **hatred** | 0 | 0 | 0 | 20 | 18 | 19 | 0 | 1 | 0 | 1 | 1 |
| **Total** | 42 | 33 | 35 | 37 | 28 | 36 | 14 | 20 | 13 | 12 | 9 |

Table 3: Statistics of the emotional words: Originals and their translations.

Then we compared the original Chinese words and their English equivalents concerning the degrees of the emotion and classified them into three groups so as to see how the emotions are transferred. It is clear that most of them are kept unchanged, but some are weakened, and there are still a few strengthened (Figure 1).

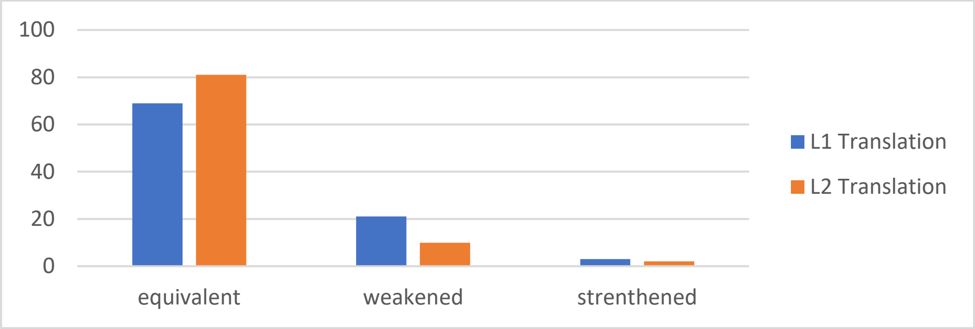


Figure 1: Change of emotions in translating emotional words.

We have also noticed several interesting cases where the diction are emotionally conflict between L1 and L2 translations. For example, “爱打麻将” is translated into “addicted to mahjong” in L1 translation, but it is “fond of playing Mahjong” in L2 translation. Going back to the original text, the reader will easily see that playing mahjong causes certain troubles. However, it seems hasty to say that the L2 translator was wrong, because within that certain context, the word “fond” reads even more sarcastically.

With this study, we are trying to explore the differences between L1 and L2 Translation. The major findings include:

(1) The emotions of most words are transferred into the target texts as they are, however, there are obvious differences between L1 and L2 translation concerning the level of diversity, equality, clarity, and intensity.

(2) L1 translations enjoy a better diversity of expressions, but tend to be more conservative when the emotions of the original words are not clear. The conservative strategies include weakening the emotion or simply avoiding translation. What’s more, L1 translations occasionally involve the personal emotions of the translators.

(3) L2 translations enjoy a much better equivalence, and when the original context is vague, translators still tries to keep or create clear emotions with their comprehension.

According to the above findings, it seems reasonable to suggest that the differences in translating emotional words be considered as new dimensions for translation quality assessment as well as translator identification. As a work in progress, however, the corpus is far from perfect, and we are observing more texts from more translators. If it is possible, more languages will be involved to testify those preliminary conclusions.

## *References*

Baker, M., & Saldanha, G. (2009). *Routledge encyclopedia of translation studies*. 2$^{nd}$ edition. London/New York: Routledge.

Campbell, S. (1998). *Translation into the second language*. London: Longman.

Graham, A. C. (1965). *Poems of the late T'ang*. Middlesex: Penguin Books.

HowNet. Retrieved from: http://www.keenage.com.

Huang, L. (2011). Translation into or out of one's native tongue- a question of directionality: A corpus-based investigation of translational styles. *Foreign Language Education*. Retrieved from: https://en.cnki.com.cn/Article_en/CJFDTotal-TEAC201102026.htm.

Jourdan, C., & Tuite, K. (2006). *Language, culture, and society: Key topics in linguistic anthropology.* New York: Cambridge University Press.

Newmark, P. A. (1988). *Textbook of translation*. New York: Prentice Hall.

Pokorn, N. K. (2005). *Challenging the traditional axioms: Translation into a non-mother tongue*. Amsterdam/Philadelphia: John Benjamins.

Wang, R., & Huang, L. (2015). A corpus-based investigation of the style between direct and inverse translations of Jia Pingwa's novels. *Foreign Languages in China*.

**Liimatta, Aatu**

***University of Helsinki***

**aatu.liimatta@helsinki.fi**

Type of Contribution: Full Paper

# Are all texts within a register comparable?

## Text length as a linguistic variable on social media

*Abstract*

In quantitative linguistic analyses, all other things being equal, texts which differ only in length are often considered comparable. When performing statistical analyses, the counts of various linguistic features can simply be normalized e.g. to a base of 1,000 words. However, very short texts tend to give anomalously high results when normalized in this way, and due to this are commonly excluded from the analysis, based on the assumption that the shorter texts within a register are "cut from the same cloth" as the longer ones and would probably contain the same distribution of features were they longer.

In most genres included in traditional corpora, texts tend to be reasonably long, so ignoring short texts is not a major problem. However, most social media postings are extremely short in comparison: their median length is typically only a few dozen words (cf. e.g. Liimatta, forthcoming; Clarke & Grieve, 2019). Since short comments constitute an overwhelming majority of social media data, they cannot be simply ignored. On the contrary, the frequencies of various linguistic features have been found to vary on Reddit in different ways as the comment length changes, particularly within the shortest comments (Liimatta, 2019b).

In this paper, I will focus on Reddit, the third-largest English-language social media platform. Reddit is made up of thousands of "subreddits", subforums which focus on various topics. In earlier studies, register differences have been found between subreddits (Liimatta, 2019a). The present study focuses on the relationship between the variation by register and the variation

by text length, investigating whether texts of all lengths within a register are really cut from the same cloth, or if different text lengths have different feature compositions *within* registers as well.

### Data

The data used in the present study has been retrieved from the Reddit Comment Corpus (Baumgartner, n.d.), a constantly updated dataset of all publicly available Reddit comments. The data used in the present study covers all comments from August 2017. After some cleaning, the dataset totals slightly under 79 million comments. This dataset was tokenized and tagged for part of speech using the Stanford CoreNLP pipeline (Manning et al., 2014). After this, each comment was analyzed for a number of lexico-grammatical features, based on Biber (1988).

The present study focuses on three subreddits, *AskHistorians*, *AskReddit*, and *INTP*, based on the finding that there is register variation between them (Liimatta 2019a, forthcoming). AskHistorians is a strictly moderated subreddit where scholars and heavily invested hobbyists answer questions about history. The answers are comprehensive and use reliable sources. Due to this, the subreddit is very academic in style, in contrast to the much more conversational style of Reddit as a whole.

In AskReddit, Reddit users ask all kinds of questions from each other and discuss various topics, such as “Do you have a famous ancestor? If so, what were they known for?” or “What is equally scary at both day and night?”.

INTP is dedicated to the INTP personality type (Introverted-iNtuitive-Thinking-Perceiving) of the Myers-Briggs Type Indicator (MBTI) personality classification system. This subreddit is meant for like-minded individuals identifying with the INTP type to discuss all kinds of topics in a casual manner.

### Method

In order to allow comparison between texts of different length, the comments are pooled together by length, so that, in the ideal case, every single text length forms its own pool. This is done separately for each subreddit to allow for comparisons. Then, the average frequency of each lexico-grammatical feature is calculated for each pool, i.e., for each text length. After this, each feature is plotted on a graph with text length on the x-axis, to see how the average frequency of the feature varies as the text length changes.

This approach lends itself particularly well to large-size social media data, as there need to be enough texts of every length for meaningful results. However, in practice, it is often useful to bin together adjacent text lengths which are less represented in the data. For example, as an overwhelming majority of social media comments are extremely short, there are enough short texts to analyze each text length separately, but texts of longer lengths often need to be binned together. In the present study, comments were binned together so that every bin contains at least 200 comments.

## Results

The results indicate that there is register variation between texts of different length even within the subreddits. This is exemplified by figures 1 and 2, which show the results for two features, first person singular pronouns and direct wh-questions. The two vertical black lines in both figures show the median text length in the full dataset, and the 90% quantile, respectively. In Figure 1, we can see the number of first-person plural pronouns for the three subreddits. First-person pronouns are an important register feature, commonly associated with registers with a highly involved style (cf. e.g. Biber, 1988; Liimatta, 2019a). At the longer end of texts, there is a clear difference between AskHistorians, where questions are answered in an academic style, and the two other subreddits, where personal anecdotes and opinions abound. However, in AskHistorians, the very shortest comments are remarkably more similar to the other two subreddits in terms of their first-person singular person use. This is because shorter comments on AskHistorians are more often not answers to questions, but the general public bringing in their own views, often asking about their understanding of the topic, driving up the number of first-person pronouns.

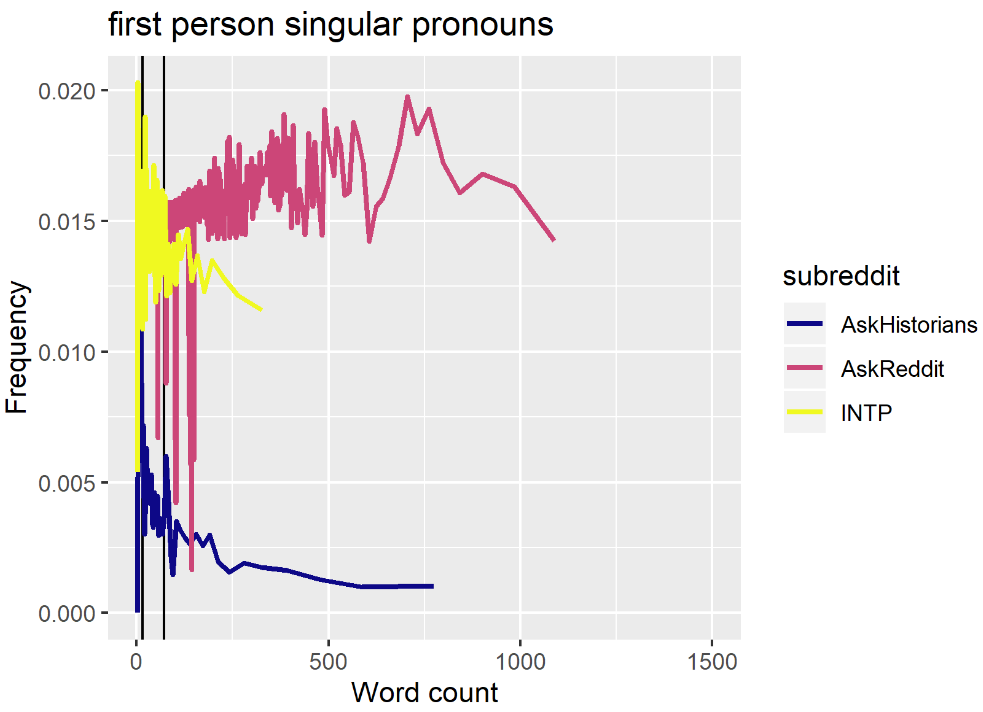


Figure 1: Frequency of first-person singular pronouns across comment lengths on the three subreddits.

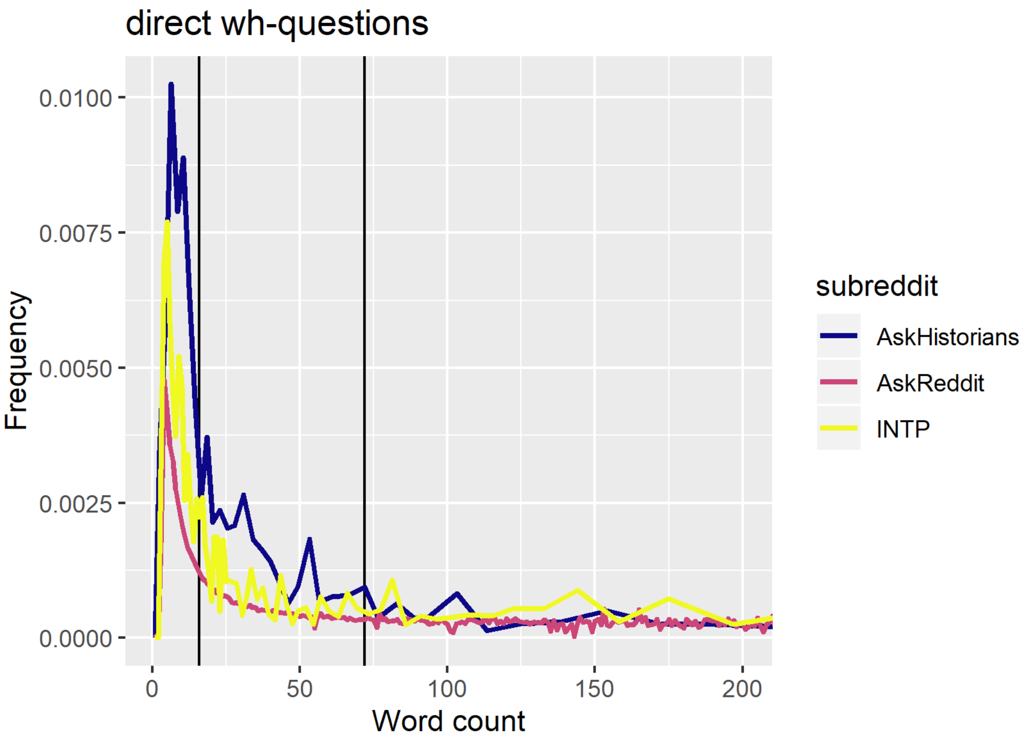


Figure 2: Frequency of direct wh-questions across comment lengths on the three subreddits.

Figure 2 shows the number of direct wh-questions across the three subreddits in shorter comments up to around 200 words. Direct wh-questions are more frequent in the shortest comments in all three subreddits. After all, it is typically not necessary to write a long text to ask a simple question. However, compared to the other two subreddits, AskHistorians has an even higher number of short questions. On the other hand, as the texts get longer, all three subreddits reach around the same frequency. This too highlights the fact that shorter comments in AskHistorians are more likely to be further questions in comparison with the other subreddits.

## Conclusion

The analysis shows not only that are there differences in register feature distributions between Reddit comments of different length, but also that the differences may differ between different subreddits. In other words, subregister variation seems to also take place between texts of different length within a register. It makes sense that certain communicative purposes would favor certain text lengths, and consequently that text length could be a register feature in itself.

## *References*

Baumgartner, J. (n.d.) Reddit Comment Corpus. Retrieved from: pushshift.io

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Clarke, I., & Grieve, J. (2019). Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLoS ONE, 14*(9), e0222062.

Liimatta, A. (2019a). Exploring register variation on Reddit: A multi-dimensional study of language use on a social media website. *Register Studies, 1*(2), 269-295.

Liimatta, A. (2019b). *Is text length a linguistic variable? Evidence from social media*. Paper presented at the 8th Biennial International Conference on the Linguistics of Contemporary English (BICLCE 8), Bamberg, Germany.

Liimatta, A. (forthcoming). Using lengthwise scaling to compare feature frequencies across text lengths on Reddit.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60).

**Mahler, Hanna**

***Albert-Ludwigs-Universität Freiburg***

**hanna.mahler@students.uni-freiburg.de**

Type of Contribution: Work-in-Progress Report

## Orthographic variation in Reddit communities

On *puppers*, *doggos*, and *good bois*

*Abstract*

The platform Reddit provides a virtual space for users to form groups around shared interests, referred to as "subreddits" (Singer et al., 2014). Within these subreddits, users often develop a topic-related slang. The focus of the study at hand are pet-themed subreddits, where users post pictures or videos of dogs and comment on them, while frequently employing a slang referred to as "doggo lingo" (Golbeck & Buntain, 2018; Punske & Butler, 2019) or "pupper talk". The slang consists of a variety of lexical creations (blends such as *puppervisor*, shortenings such as *pup*, derivations such as *treatos*), irregular grammatical constructions (irregular morphology such as *goodest*, constructions such as *do a dilemma*), as well as non-standard orthography (omission of letters as in <fren>, substitution of letters as in <boi> or <snoot>). The slang incorporates elements from other online slangs, such as LOLspeak (Bury & Wojtaszek, 2017), and has structural similarities to speech during first language acquisition.

Considering the usage of this slang as an instance of stylistic variation, my study addresses the following research question: Which factors influence the presence of community-specific features in the posts and comments of pet-themed communities on Reddit? As a framework I take up Bell's (1984) distinction of audience and non-audience factors but propose new audience roles for the anonymous Reddit environment: the *formal addressee* (the author of the preceding comment), the *active audience* (users engaging in up- and down-voting of comments), and the *passive audience* (people passively consuming content).

To answer the research question, quantitative and qualitative methods were combined. First, I completed a pilot study using online ethnography (Androutsopoulos, 2008), conducting eight written interviews with community moderators. This was followed by the compilation of a corpus, which contains 356 posts and 4472 responding comments, collected from eight subreddits varying in size and attitude towards the slang. Posts and comments were collected manually on a weekly basis, selecting the five most popular posts from each subreddit, accompanied by a fixed number of comments. The following meta-data were collected for each comment: subreddit, author, date, text, comment karma (an indication of community appreciation), level, which post and comment it responds to, presence of slang in previous textual unit, and number of words. The quotes below (1-4) provide an example of comments on the three different levels using the slang.

(1) This might qualify as the bestest boyyy ever…

(2) Betterest Boy!!!

(3) Everyone thinks they have the bestest boy ever; and everyone is right

(4) And if they don't they have the bestest girl!

A multiple logistic regression model was fitted to the data to identify predictors influencing the presence or absence of the slang. The selection of predictors for the model was based on existing literature and the findings of the qualitative pilot study. Potential predictors influencing the style are the subreddit, the type and "cuteness" of the accompanying visual material (approximated through age of the pet, visibility of the face, and anthropoidness), the usage of slang within the previous comment, and the position of the text within the comment structure (level 1, 2, or 3).

The regression model (d.f. 17, Pr <0.0001, $R^2$ 0.057, C 0.647) revealed that only the level of the comment and the style within the previous textual unit have a significant impact on the probability of pupper talk occurring. Comments on level 2 and 3 have a smaller likelihood of using the slang (Coef -0.395, <0.0001, and -0.384, 0.0005), which is probably related to the topic drifting away from the pet depicted. This emphasizes firstly the importance of topic as a non-audience factor and secondly the topic-boundedness of the slang. On the other hand, previous usage of the slang leads to a higher likelihood of the slang occurring (Coef +0.754, <0.0001). This indicates that the style employed by the formal addressee has a major impact on users' stylistic choices. One could also speculate that by adapting to the style of the formal addressee, users want to capture the same appreciation by the active audience as the previous author, which would position the active audience as the most influential audience role. This assumption is supported by comments made by informants during the pilot study. All other factors, including those explicitly named by the moderators, did not reach statistical significance.

In my study I also provide a detailed discussion of the applicability of the two terms *virtual community* (Herring, 2004) and *community of practice* (Meyerhoff, 2004) to the subreddits under investigation. While every subreddit can be classified as a community of practice due to their distinct enterprises (despite a considerable overlap in shared repertoire), not all of them fulfil the criteria for virtual communities. All subreddits have the structural prerequisites to develop a virtual community, such as explicit rules, access to previous content, and moderators, but some subreddits use additional tools to foster a sense of community, such as explicitly endorsing the use of pupper talk and joint language play. This finding emphasizes that these terms can only be applied after careful consideration.

In my paper I also draw attention to several methodological challenges for the quantitative study of stylistic variation in the online environment. The most important one concerns the measurement of stylistic variation for any textual unit. Compared to a binary variable (presence or absence of the slang), a numeric measurement (such as a ratio of the number of features in relation to the number of words) would provide more insights but was proven to be unfeasible for the data at hand. The main reason lies within the overall shortness of the comments (mean length: 12.4 words), which disproportionally favours ratios such as 0.5, making a linear regression model impossible. A further issue resulted from the sampling method chosen. Through sampling by time, 83 percent of the comments collected did not contain any slang and received a ratio of zero, which further complicates the applicability of linear regression for the data at hand.

My study therefore contributes to understanding stylistic variation and weighing audience and non-audience factors in the online environment by combining quantitative and qualitative methods. Furthermore, it illustrates the importance of slang and humour for the creation of virtual communities (e.g. North, 2007). By providing a complete description of the slang "pupper talk" the study also contributes to the ongoing exploration of language use on Reddit. Additionally, my study is one of the first to take into account the multimodal environment of the platform (Herring, 2015) and its potential influence on stylistic variation. The findings of the analysis will likely be transferable to stylistic choices on other subreddits and on similar platforms.

## *References*

Androutsopoulos, J. (2008). Potentials and limitations of discourse-centred online ethnography. *Language@Internet* 5, 1-20.

Bell, A. (1984). Language style as audience design. *Language in Society, 13*(2), 145-204.

Bury, B., & Wojtaszek, A. (2017). Linguistic regularities of LOLspeak. *Sino-US English Teaching, 14*(1), 30-41.

Golbeck, J., & Buntain, C. (2018). This Paper is About Lexical Propagation on Twitter: H*ckin Smart. 12/10. Would Accept! In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 587-590). Barcelona: IEEE.

Herring, S. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S. Barab, R. Kling, & J. Gray (Eds.), *Designing for virtual communities in the service of learning* (pp. 338-376). New York: Cambridge University Press.

Herring, S. (2015). New frontiers in interactive multimodal communication. In A. Georgakopoulou & T. Spilioti (Eds.), *The Routledge Handbook of Language and Digital Communication* (pp. 398-402). London: Routledge.

Meyerhoff, M. (2004). Communities of practice. In J. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (pp. 526-548). Malden: Blackwell.

North, S. (2007). 'The voices, the voices': Creativity in online conversation. *Applied Linguistics, 28*(4), 538-555.

Punske, J., & Butler, E. (2019). Do me a syntax: Doggo memes, language games and the internal structure of English. *Ampersand, 6*, 1-9.

Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., & Strohmaier, M. (2014). Evolution of Reddit: From the front page of the internet to a self-referential community? In *International World Wide Web Conference Committee* (IW3C2) (pp. 517-522). Seoul.

**Malá, Markéta**

***Charles University***

**marketa.mala@ff.cuni.cz**

**Raušová, Veronika**

***Charles University***

**veronika.rausova@gmail.com**

Type of Contribution: Poster Presentation

## Phraseology in learner academic writing

The case of *it*-bundles

*Abstract*

The idiomaticity of discourse depends to a large extent on pragmatically appropriate use of recurrent multi-word expressions. Such phraseological units and their functions were shown to be register-dependent, constituting “the preferred way of saying things in a particular discourse” (Gledhill, 2000, p. 202). For novice academic writers, phraseology is therefore a means of indicating that they belong to the discourse community (Hyland, 2008). At the same time, L2 novice writers face the foreign language challenge, and it is phraseology that distinguishes between native speakers and advanced L2 learners (Ebeling & Hasselgård, 2015). We explore the ways and extent to which the two challenges, the foreign language challenge (EFL) and the “academic” challenge (EAP), affect Czech novice writers’ use of phraseological patterns.

Our study combines contrastive analysis and learner corpus research. It explores multi-word patterns comprising the word-form *it* in English L2 academic texts written by Czech university students in comparison with English L1 novice and expert writing. Grammatical, or “small” words have been shown to be “crucial to textual meaning” (Hunston, 2008, p. 272). We have therefore selected one of the most salient grammatical keywords for the English L2 texts (compared to a reference corpus of English academic papers) as our starting

point. Two dimensions of contrast are explored: novice-expert (English texts written by novice academic writers, L1 and L2, are compared with those published in academic journals), and native-learner English. The analysis draws on three corpora comprising English academic texts (essays and published papers) dealing with English literature (see Table 1).

| **corpus** | **L2-novice** | **L1-novice** | **L1-expert** |
|---|---|---|---|
| **source** | Charles University, Prague, Faculty of Arts, English Studies, BA | BAWE, Arts and Humanities - English | academic journals |
| **time** | 2016-19 | 2004-17 | 1978-2014 |
| **register** | students' essays | students' essays | academic papers |
| **size: tokens (approx.)** | 106 600 | 226 300 | 235 000 |
| **number of texts** | 48 | 89 | 34 |

Table 1: Composition of the corpora.

3-grams comprising *it* were identified in the three corpora, and their frequencies were compared. We singled out *it*-grams which were over- or underused significantly by both groups of novice writers or by English L2 writers in comparison with L1 expert writers (i.e. the norm, cf. Römer, 2009, p. 89). The patterns including these *it*-grams and their functions were explored in detail. We assume that patterns over/underused by both groups of apprentice academic writers signal the impact of the EAP challenge: “The fact that native and non-native apprentice academic writers lack very similar sets of expert academic English phraseological items in their papers indicates that both groups of students may need similar training or help with their academic writing on their way to becoming more proficient writers" (Römer, 2009, p. 99). Where the patterns used by Czech students differ from those used by both groups of English L1 writers, EFL problems may be detected.

The results demonstrate that phraseological differences in the use of *it*-patterns between Czech novice academic writers and English expert writers are indeed due to both types of challenge the students have to face, EAP and EFL, with the academic expertise being perhaps a more prominent one at this proficiency level. Still, the EFL challenge manifests itself in the writing of the Czech students especially in their underuse of more complex patterns needed for information packaging (e.g. cleft constructions (ex. 1), postposed content clauses, shell nouns, *as it does*), and in their choice of stance patterns (overuse of *it is obvious*, *it is true* (ex. 2)). In both areas, the impact of Czech may be suspected (*je pravda, že* - "it is true that"; different means of indicating information structure) and should be further explored.

(1) *What makes the tragedy of Othello so shocking and painful <u>is that it</u> engages its audience* […] (L1-expert)

(2) *<u>It is obvious</u> from the context that the china in question is not a real piece of pottery but carries a double, sexual subtext*. (L2-novice)

The EAP challenge seems to affect the students' (both English L1 and L2) use of *it*-patterns more noticeably, involving stylistic missteps (e.g. *when it comes (to), it is interesting* (ex. 3)) and underusing conventional hedging strategies (e.g. impersonal patterns expressing stance - *it could/might be*). Where several functionally equivalent means are available, the apprentice academic writers may display different preferences than experienced writers (e.g. *although, though, even though*, ex. 4, cf. also Mbodj & Crossley, 2020).

(3) *<u>It is interesting</u> to note that both these terms are used in reference to money* […] (L1-novice)

(4) *<u>Even though it</u> is true that Mosca is very compliant to do anything his master Volpone sets his mind to, nonetheless Mosca is the one who comes up with most of the schemes and lies*. (L2-novice)

It has been argued that simply exposing students to successful expert academic writing does not effectively improve their own (Dontcheva-Navratilova, 2012). Therefore, based on our findings, it would be helpful to familiarize Czech students with various hedging strategies and patterns employed to express them, and focus in particular on stylistic appropriacy. We hope to have shown that linguistic training, focussing of information structure and raising awareness of the differences between English and Czech, should not be neglected at this level of proficiency either.

## *References*

Ebeling, S., & Hasselgård, H. (2015). Learners' and native speakers' use of recurrent word-combinations across disciplines. *Bergen Language and Linguistics Studies, 6*, 87-106.

Dontcheva-Navratilova, O. (2012). Lexical bundles in academic texts by non-native speakers. *Brno Studies in English, 38*(2), 37-58.

Gledhill, C. (2000). *Collocations in science writing*. Tübingen: Gunter Narr Verlag.

Hunston, S. (2008). Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics, 13*(3), 271-295.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21.

Mbodj N. B., & Crossley, S. (2020). Students' use of lexical bundles: Exploring the discipline and writing experience interface. In Römer, U., V. Cortes & E. Friginal (Eds.), *Advances in corpus-based research on academic writing: Effects of discipline, register, and writer expertise* (pp. 115-135). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Römer, U. (2009). English in academia: Does nativeness matter? *Anglistik: International Journal of English Studies*, *20*(2), 89-100.

**Malá, Lucie**

***Charles University***

**luckasmile@yahoo.co.uk**

Type of Contribution: Work-in-Progress Report

## Phraseology of mathematical texts

Constructional approach

*Abstract*

A growing body of research has focused on the differences between the academic writing of individual disciplines (e.g. Hyland, 2004, 2008; Gray, 2015). However, mathematical texts in general, and specifically mathematical research articles, are underrepresented in research into phraseology of scientific texts. The present study focuses on the distributional phraseology of mathematical texts. The aim is the identification and constructional description of key elements of mathematical research articles, which can be seen as the basic building blocks of this genre.

Although the goals of this study are descriptive rather than theoretical, the choice of the framework of construction grammar is well justified. It has been already shown that construction grammar can be used in characterizing specific disciplinary discourses (Groom, 2019; Hiltunen, 2010). Its main benefits are a concise and unified description of the key phraseological units across all language dimensions, which is of particular importance for any possible future pedagogical outcomes of the present research. Moreover, construction grammar takes into account not only the description of the individual units, but also their mutual relationships. It is therefore possible to organize the constructions into a network based in their shared features.

In accord with the general principles of construction grammar, the study is corpus-driven, relying on inductive methods for identification of the key elements. The corpus used (870,885 words) has been compiled for this purpose, and includes mathematical research papers from three fields, namely

algebra, mathematical analysis, and probability & statistics. The main method applied is keyword extraction. The corpus of the target texts is compared with a reference corpus of published academic papers across 28 disciplines, i.e. the Corpus of Academic Journal Articles (Kosem, 2010). As this corpus is available in the Sketch Engine software, the Sketch Engine keyword extraction was used to retrieve the keywords. To make the results more reliable, this automatic extraction was combined with calculating confidence intervals for each of the first 10,000 words suggested as keywords by the software. By comparing the intervals in the focus corpus with those in the reference corpus for each word, we made sure that not only is the word significantly more frequent in our corpus, but that with 95% confidence it is more significant in any other similar corpus representative of the same text type. The thus obtained keyword list was then manually analyzed. Following the methodology promoted by Groom (e.g. 2010) or Gledhill (2000), the focus is on grammatical keywords, which are expected to offer better coverage of the text and be more suitable as starting points for a phraseological analysis.

15 grammatical keywords have been extracted, namely *let, we, then, if, where, hence, every, since, now, any, moreover, whenever, therein, whence, otherwise*. These were compared with the lists of grammatical keywords characteristic of journal papers from history and literature compiled by Groom (2007). None of the extracted keywords was found in either of the two lists. This signals that the extracted grammatical keywords are most likely specific to mathematics research papers.

Concordance lines for each of the keywords are analyzed in detail in order to identify constructions it participates in. This study works with the usage-based definition of a construction, which suggests that any pairing of form and meaning can be recognized and stored as a construction “even if they are fully predictable as long as they occur with sufficient frequency” (Goldberg, 2006, p.5). It has been found that each of the words under investigation participates in one construction at least. The word *let*, for instance, participates in four different constructions, provisionally named LET US DO SOMETHING, PRESENTATIVE LET, ASSUMPTIONS, and WE LET. While some of these constructions have been to some degree already suggested in literature (see Cunningham, 2017; Swales et al., 1998), they have not been described with the amount of precision and detail relevant to their use in mathematical research papers brought by the present study before.

The ultimate aim of this research is a detailed description of all the constructions related to the keywords. These constructions will be organized into a network representing a local constructicon of mathematical research papers.

## *References*

Cunningham, K. J. (2017). A phraseological exploration of recent mathematics research articles through key phrase frames. *Journal of English for Academic Purposes*, *25*, 71-83.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, *19*(2), 115-135.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Gray, B. (2015). *Linguistic variation in research articles*. Amsterdam: John Benjamins.

Groom, N. (2007). *Phraseology and epistemology in humanities writing: A corpus-driven study*. PhD Thesis.

Groom, N. (2010). Closed-class keywords and corpus-driven discourse analysis. In M. Bondi & M. Scott (Eds.), *Keyness in text* (pp. 59-78). Amsterdam: John Benjamins.

Groom, N. (2019). Construction Grammar and the corpus-based analysis of discourses: The case of the WAY IN WHICH construction. *International Journal of Corpus Linguistics*, *24*(3), 291-323.

Hiltunen, T. (2010). *Grammar and disciplinary culture: A corpus-based study*. PhD Thesis. University of Helsinki.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. The University of Michigan Press.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4-21.

Kosem, I. (2010). CAJA: Corpus of Academic Journal Articles | Sketch Engine. Retrieved from https://www.sketchengine.eu/corpus-of-academic-journal-articles-caja/.

Sketch Engine. Retrieved from: http://www.sketchengine.eu.

Swales, J., Ahmad, U., Chang, Y.-Y., Chavez, D., Dressen-Hammouda, D., & Seymour, R. (1998). ‘Consider This...’: The role of imperatives in scholarly writing. *Applied Linguistics*, *19*, 97-121.

**Marcus, Imogen**

**Edge Hill University**

**marcusi@edgehill.ac.uk**

**Maden-Weinberger, Ursula**

**Edge Hill University**

**weinberu@edgehill.ac.uk**

Type of Contribution: Full Paper

# From manuscripts to messaging

## Historical perspectives on orality in CMC

*Abstract*

The main dimensions of innovation in digital written language fall into three themes according to (Androutsopoulos, 2007): orality, compensation, economy. We are focusing on the first theme: (conceptual) orality, which refers to linguistic features characteristic of casual spoken language occurring in written discourse (cf., e.g. Baron, 1984; Schmitz, 2001). Computer-mediated communication (henceforth CMC) has been conceptualized as a 'blend or hybrid of written and spoken aspects of language' (Androutsopoulos, 2011, p. 5). The link between coordination and prototypically spoken communication on the one hand, and subordination and prototypically written communication on the other has been established by a number of scholars (cf. e.g. Jahandarie, 1999, p. 144-145; Biber et al., 1999).

The current paper therefore focuses on three clause-level coordinating conjunctions, AND, BUT and OR. It asks: what are the frequencies of clause-level AND, BUT and OR in four different sub-corpora: email, instant messaging (henceforth IM), sermons and statutes? Sermons are "speech-purposed" (Culpeper and Kytö, 2010, p. 18) so sit at the 'spoken' end of a conceptual speech-writing continuum. Conversely, statutes are considered to be "writing-based and purposed" (ibid.) and sit at the 'written' end of the continuum. What

does a comparison of frequencies of AND, BUT and OR in these four text types therefore tell us about where email and IM sit on the conceptual speech-writing continuum? How do the ‘connective profiles’ of email and IM compare to the profiles of sermons and statutes? Can any frequency changes be observed over the 16th-21st centuries in the sermon and statute data?

Previous research that has compared CMC to earlier writing (cf. e.g. Elspaß, 2002; Baron, 2008; Bergs, 2009; Shortis, 2009; Tagg & Evans, forthcoming, 2020), has concluded that observations about the novelty of digital writing tend to be exaggerated and/or lack historical depth. The current paper takes a quantitative, diachronic corpus linguistics approach (cf. Hilpert & Gries, 2016) to the use of coordinators in digital writing in the context of sermons and statutes, and aims to add to this existing body of diachronic research.

We are broadly following the ‘connective profiling’ methodology first proposed by Kohnen (2007). Connective profiling involves investigating “the distribution, frequency and proportion of the major clause-connecting coordinators and subordinators” (Kohnen, 2007, 289) in different text types. The rationale behind the methodology is that it provides an alternative approach to most studies of connectives in the history of English, which “tend to focus on the development of individual items and their position in the language system” (2007, p. 289). Connective profiles provide a more comprehensive picture and “may highlight the relation of clause-connective elements to the orality and literacy of texts” (2007: 289), thereby helping us to better understand language change. Kohnen (2007) focuses on sermons and statutes; the current paper broadens the comparison to include email and IM. The final published version of this paper will also include the analysis of a correspondence (letters) sub-corpus as part of the OWT (Orality in Written Texts) corpus, which we compiled for this study. OWT consists of five text types: statutes, sermons, letters, emails and IM. For statutes, sermons and letters, we have data stretching from the 15th to the 21st century. Our email and IM sub-corpora date from the 21st century.

Our data analysis in this presentation focuses on the coordinators AND, BUT and OR. We have found these lexical items to be fairly evenly dispersed within all sub-corpora. Sermons (‘speech purposed’) display higher frequencies of the coordinating conjunctions AND and BUT compared to statutes (‘writing based/purposed’), which is what we would expect, given the link between coordination and spoken communication. Clause-level OR bucks the trend, as this lexical item has higher frequencies in statutes than in sermons. However, the overall frequencies of OR are relatively low across all sub-corpora compared to AND and BUT. Broadly, these findings are in-keeping with those of Kohnen (2007).

The relative frequencies of AND, BUT and OR in both email and IM are higher than their relative frequencies in statutes and lower than their relative frequencies in sermons, meaning both varieties of CMC fall in the middle. This finding provides more evidence to support the conceptualization of CMC as a blend of spoken and written language. Differences in frequencies have been found to be statistically significant for AND, BUT and OR between IM and both sermons and statutes. For emails, frequency differences are statistically significant between emails and sermons, but not between emails and statutes. This positions email closer to the written end of the continuum where the statutes are situated, which could be related to the fact that email is more asynchronous than IM.

## *References*

Androutsopoulos, J. (2007). Neue Medien - neue Schriftlichkeit? *Mitteilungen des Deutschen Germanistenverbandes, 1*(7), 72-97.

Androutsopoulos, J. (2011). Language change and digital media: a review of conceptions and evidence. In T. Kristiansen & N. Coupland (Eds.), *Standard languages and language standards in a changing Europe* (pp. 145-161). Oslo: Novus.

Baron, N. S. (1984). Computer mediated communication as a force in language change. *Visible Language, 18*(2), 118-141.

Baron, N. S. (2008). *Always on: Language in an online and mobile world*. Oxford: Oxford University Press.

Bergs, A. T. (2009). Just the same old story? The linguistics of text messaging and its cultural repercussions. In C. Rowe & E. L. Wyss (Eds.), *Language and new media* (pp. 55-73). Cresskill, NJ: Hampton.

Biber, D, Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Culpeper, J., & Kytö, M. (2010). *Early Modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.

Elspaß, S. (2002). Alter Wein und neue Schläuche? Briefe der Wende zum 20. Jahrhundert und Texte der neuen Medien - ein Vergleich. *Osnabrücker Beiträge zur Sprachtheorie, 64,* 7-32.

Hilpert, M., & Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. In M. Kytö & P. Pahta (Eds.), *The Cambridge handbook of English historical linguistics* (pp. 36-53). Cambridge: Cambridge University Press.

Jahandarie, K. (1999). *Spoken and written discourse: A multidisciplinary perspective*. Stamford: Ablex.

Kohnen, T. (2007). 'Connective profiles' in the history of English texts: Aspects of orality and literacy. In U. Lenker & A. Meurmann-Solin (Eds.), *Connectives in the History of English* (pp. 289-308). Amsterdam/Philadelphia: John Benjamins.

Schmitz, U. (2001). Auswirkungen elektronischer Medien und neuer Kommunikationstechniken auf das Sprachverhalten von Individuum und Gesellschaft. In W. Besch, A. Betten, O. Reichmann, & S. Sonderegger (Eds.), *Sprachgeschichte* (pp. 2168-2175). 2nd edition. Vol. 2. Berlin/New York: de Gruyter.

Shortis, T. (2009). Revoicing Txt: Spelling, vernacular orthography and 'unregimented writing'. In S. Wheeler (Ed.), *Connected minds, emerging cultures: Cybercultures in online learning* (pp. 225-246). Charlotte, NC: IAP.

Tagg, C., & Evans, M. (Forthcoming, 2020). Spelling in context: A transhistorical pragmatic perspective on orthographic practices in English. In C. Tagg & M. Evans (Eds.), *Message and medium: English language practices across old and new media* (pp. 55-79). Berlin: Mouton De Gruyter.

**Notohara, Yoshiyuki**

***Doshisha University***

**ynotohar@mail.doshisha.ac.jp**

Type of Contribution: Full Paper

# Exploring the metaphorical and image-schematic grounding of indirect speech acts realization in processes and force-dynamic schemata constructions

A corpus pragmatics approach

*Abstract*

This study explores indirect speech acts realization in Radden & Dirven's (2007) five canonical constructions related to processes and force-dynamic schemata constructions through the Spoken BNC2014 corpus study. It also interprets the relationships between constructional meanings (e.g., Transfer) and indirect speech acts realization (e.g., suggesting) and theoretically proposes metaphorical and image-schematic grounding of indirect speech acts (e.g., Hernandez & Mendoza, 2002) in canonical constructions.

So far, in pragmatics, the relationships between indirect speech acts realization and utterance forms have been explored and discussed in terms of conventionality (e.g., Ruytenbeek, Ostaschchenko, & Kissine, 2017). Quite recently, corpus pragmatics has been exploring indirect speech acts realization in several utterance forms such as formulaic sequences, clause types and implicit causality through both bottom-up and top-down approaches (e.g., Rühlemann, 2019). However, there is little related research on canonical constructions. Although speech acts realization in frequent material-based schemata (e.g., States/SVC) and psychological-based schemata (e.g., Emotion/SVO) constructions have been explored so far, less frequent two material-based schemata (e.g., Object-motion/SV, Processes/SVC) and three force-dynamic schemata (e.g., Action/SVO, Caused-motion/SVO, Transfer/SVO) constructions have not been studied yet. Therefore, two

research questions (RQs) are addressed here: RQ1: What indirect speech acts are realized in less frequent five canonical constructions?; RQ2: What kinds of relationships are there between indirect speech acts realization and the constructions?

(1) Less frequent five canonical constructions
a. *I'm getting better.* (Processes/SVC)
b. *The prize goes to a child.* (Object-motion/SV)
c. *I'll make some tea.* (Action/SVO)
d. *He put the platter on the floor.* (Caused-motion/SVO)
e. *She gave it to the mouth.* (Transfer/SVO)

The procedure for the study was as follows: (1) 1,000 usages of each target verb (e.g., give (Transfer/SVO)) were randomly selected from the Spoken BNC2014 corpus (approximately 10 million words); (2) indirect speech acts in exemplars were respectively interpreted and coded referring to van Ek & Trim's (1998) six language functions subcategorizing 132 indirect speech acts (i.e., (a) imparting and seeking factual information, (b) expressing and finding out attitudes, (c) getting things done (suasion), (d) socializing, (e) structuring discourse, and (f) communication repair); (3) the relationships between five canonical constructions and emerging indirect speech acts were statistically confirmed through the correspondence analysis; (4) the relationships between constructional meanings (e.g., Transfer) and indirect speech acts realization (e.g., suggesting) were discussed.

As a result, it was found that indirect speech acts related to language functions (a) and (b) were most often seen in five canonical constructions (χ2(100)=4503.742, *p*= .000 95%CI [.000-.000], Cramer's V = . 475, *p*= .000, 95%CI [.000-.000]). More specifically, (1) All five canonical constructions have a common indirect speech act, namely, reporting; (2) Processes/SVC tended to be related to negative indirect speech acts (e.g., expressing dissatisfaction, fear, and worry); and (3) Action/SVO, Caused-motion/SVO, Transfer/SVO, Object-motion/SVO tended to be related to interactive indirect speech acts (e.g., suggesting, encouraging, declining). Finally, based on the results, the metaphorical and image-schematic grounding of indirect speech acts realization in five constructions in the current study are theoretically discussed.

(2) Common indirect speech act (Reporting)

a. *it just goes everywhere.* (Object-motion/SV, Reporting)

b. *it was getting a bit warm.* (Processes/SVC, Reporting)

c. *they are making a film.* (Action/SVO, Reporting)

d. *I put a picture on Facebook.* (Caused-motion/SVO, Reporting)

e. *they gave it to the man.* (Transfer/SVO, Reporting)

(3) Negative indirect speech acts (Processes/SVC)

a. *oh no you get dirty definitely.* (Processes/SVC, Dissatisfaction)

b. *I'm getting confused here.* (Processes/SVC, Fear)

(4) Interactive indirect speech acts (suggesting, encouraging, declining)

c. *shall we put a net around it or something?* (Caused-motion/SVO, Suggesting)

b. *you can make whatever you like.* (Action/SVO, Encouraging)

c. *I'll give them to someone else.* (Transfer/SVO, Declining)

## *References*

Hernandez, L.P., & de Mendoza, F.J.R. (2002). Grounding, semantic motivation, and conceptual interaction in indirect directive speech acts. *Journal of Pragmatics, 34*(3), 259-284.

Radden, G., & Dirven, R. (2007). *Cognitive English grammar.* Amsterdam: John Benjamins.

Rühlemann, C. (2019). *Corpus linguistics for pragmatics: A guide for research.* Abingdon: Routledge.

Ruytenbeek, N., Ostaschchenko, E., & Kissine, M. (2017). Indirect request processing, sentence types and illocutionary forces. *Journal of Pragmatics, 119*, 46-62.

van Ek, J. A., & Trim, J. (1998). *Threshold 1990: The revised and corrected edition.* Cambridge: Cambridge University Press.

**Põldvere, Nele**

***Lund University***

**nele.poldvere@englund.lu.se**

**Johansson, Victoria**

***Lund University***

**victoria.johansson@ling.lu.se**

**Paradis, Carita**

***Lund University***

**carita.paradis@englund.lu.se**

Type of Contribution: Full Paper

## The new London-Lund Corpus 2

Key methodological challenges and innovations

*Abstract*

A few years ago, at ICAME, we presented a work-in-progress paper on the very first stages of compiling the new London-Lund Corpus 2 (LLC-2). This year, we are in the fortunate position of being able to present the final corpus and to reflect on the whole compilation process. We do this by describing and critically examining three main methodological challenges that we encountered during the compilation of the corpus.

LLC-2 is a half-a-million-word corpus of contemporary spoken British English, collected 2014-2019. Its size and design are comparable to that of the world’s first machine-readable spoken corpus, the London-Lund Corpus (LLC-1) with data mainly from the 1960s. With LLC-2, we now not only have a new spoken corpus, but also a corpus that gives researchers the opportunity to make principled diachronic comparisons of speech over the past 50 years and to detect change in communicative behavior among speakers. Through a

critical discussion of the methodological challenges of compiling LLC-2, we propose solutions to overcoming the challenges, often in innovative ways, as well as facilitate more informed uses of the corpus in the future.

The compilation of LLC-2 included a number of different stages: the design of the corpus, ethical and legal considerations, recruiting participants, data collection, transcription of the recordings, markup and annotation procedures, and finally making the corpus accessible to the research community. Each stage presented its own methodological challenges. The first challenge discussed in this paper concerns the design of LLC-2. More specifically, the challenge was to strike a balance between LLC-2 as a representative collection of contemporary spoken English and its usefulness for diachronic comparisons with speech in LLC-1. According to Leech (2007), representativeness and comparability are inherently incompatible in corpus compilation since improvements in one lead to loss in the other. To achieve a sufficiently satisfactory balance between the two concepts, LLC-2 contains the same speech situations as LLC-1, i.e., primarily dialogue (in particular everyday face-to-face conversation) but also monologue; however, the specific recordings added to LLC-2 also reflect the technological advances of the past few decades, particularly with respect to speech situations such as telephone calls (e.g., Skype), and broadcast discussions and interviews (e.g., podcasts). As a result, LLC-1 and LLC-2 are comparable in the sense that they differ from each other in terms of only one parameter, the parameter of time, but, on its own, LLC-2 is also representative of the different kinds of communication channels used in the 21st century.

Another challenge concerned data collection. The distribution of key demographic categories such as age and gender in LLC-2 is relatively even. For example, there is only slight skewness in the data towards speakers aged between 35 and 59 years old. Achieving this balance was, however, not easy because of the large number of speakers needed to complete the corpus, the time constraints of the project, and the nature of the task at hand (i.e., to record a 30-minute conversation with other people). Therefore, our approach to data collection was largely opportunistic, meaning that we did not actively seek out recordings from certain groups of speakers, but instead accepted recordings from all speakers willing to be recorded for inclusion in the corpus. However, when it became clear at later stages of the data collection that some demographic categories were heavily skewed towards certain groups of speakers, efforts were made to reduce the bias to the extent possible. Moreover, the inclusion of online recordings in LLC-2 such as broadcast discussions and interviews allowed us to target specific demographic groups more easily, largely thanks to the wealth of data found online.

The third challenge discussed in this paper concerned transcriptions and markup. More specifically, it was important to develop a transcription and markup scheme that at the same time was (i) economical, (ii) useful for a wide range of areas in corpus linguistics, and (iii) compatible with modern corpus linguistic and natural language processing tools. Therefore, the transcriptions in LLC-2 are orthographic and involve a manual transcription of words together with markups of basic spoken features. The transcription and markup scheme was kept as simple as possible, largely in consideration of the workload of the transcribers. Moreover, the scheme draws heavily on the guidelines of the International Corpus of English (ICE), which have proved useful for investigations of spoken English in a wide range of areas in corpus linguistics such as lexicology, morphosyntax and discourse analysis. In contrast to ICE, however, the transcriptions in LLC-2 are based on the standardized markup language XML, which is compatible with many of the well-known text processing tools currently used by corpus linguists such as AntConc and Wordsmith Tools. A feature of LLC-2 that is entirely new is that the transcriptions are time-aligned with the audio files, and they are released together in order to allow users to extend the orthographic transcriptions relative to their own research interests such as for prosodic research (see Põldvere et al., 2020 for the main challenges encountered in the public release of the LLC-2 audio material such as the anonymization of the audio files).

In sum, the three main challenges and the innovations related to the compilation of LLC-2 concerned (i) the design of the corpus, (ii) data collection, and (iii) developing an effective and flexible transcription and markup scheme. As a critique of the solutions proposed in each stage, this paper serves as an example of how researchers may address recurring issues of spoken corpus development and analysis in the future.

## *References*

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 133-149). Amsterdam/New York: Rodopi.

Põldvere, N., Frid, J., Johansson, V., & Paradis, C. (2020). Challenges of releasing audio material for spoken data: The case of the London-Lund Corpus 2. Manuscript submitted for publication.

**Rautionaho, Paula**

***University of Eastern Finland***

**paula.rautionaho@uef.fi**

Type of Contribution: Full Paper

# *You must be knowing this*

## The stative progressive in World Englishes

### *Abstract*

Wider use of the stative progressive (*be* V*ing*) is commonly attributed to English as a second language (ESL) varieties; progressives referring to non-temporary states are seen as characteristic of Indian English (IndE), in particular (e.g. Sharma, 2009; Hundt & Vogel, 2011). However, studies making this claim often lack conclusive evidence as they almost exclusively focus on the progressive alone and thus fail to account for the fact that the overall frequency of a particular lemma may differ from one variety to another (see Aarts et al., 2013: 19). This study remedies the methodological problem by focusing on the progressive vs. non-progressive alternation in the use of 12 stative verbs, chosen based on earlier findings in Deshors and Rautionaho (2018), and Rautionaho and Fuchs (2020). In addition to calculating the proportion of progressive usage, more advanced statistical methods, namely multiple distinctive collexeme analysis (MDCA), conditional random forest analysis and conditional inference tree (ctree) analysis, are used to thoroughly assess the claim that ESLs are characterized by wider use of the progressive. "Wider use" is understood as both more *frequent* use and *extended* use to contexts where standardized varieties would not use the progressive. The research questions are as follows:

(1) Are stative progressives more frequent in non-native varieties, compared to native varieties?

(2) How do the six varieties differ with respect to their collostructional preferences?

(3) Which linguistic factors contribute to the patterns of the progressive vs. simple alternation in combination with stative verbs?

(4) To what extent do individual English varieties yield different alternation patterns?

This study focuses on three native varieties, British (BrE), Irish English (IrE) and New Zealand (NZE) Englishes, and three ESLs, Indian (IndE), Singaporean (SgE) and Hong Kong (HKE) Englishes. All occurrences of 12 stative verbs (*be*+adj, *exist*, *have*+noun, *hear*, *know*, live, *look*+adj, *love*, *reside*, stay, *understand*, *want*) were extracted from the relevant components of the *International Corpus of English* (ICE), and the analysis was restricted to a variable context where a progressive could potentially occur (excluding, for example, *there*-existentials, imperatives and idiomatic expressions such as *you know*; see e.g. Rautionaho & Fuchs, 2020). Once dynamic uses of the stative progressives were manually excluded (e.g. *having dinner/problems*), the final database contains 16,060 tokens. The data was then annotated for a number of variables; ASPECT is the dependent variable and the predictor variables include, for instance, FORM of the verb phrase (present, past, modal, perfect), ANIMACY of the grammatical subject (animate, human, inanimate), semantic GROUP (affective, cognitive, perception, relational, stance), and variety TYPE (ENL, ESL).

The results indicate very clearly that instead of ESLs showing wider use of stative progressives, higher frequency and extended uses are found in IndE. The proportional frequency of stative progressives (calculated as a proportion of progressive instances of all instances of a particular lemma) in IndE is 6.5%, while in the three ENLs it is c. 3.7% and in SgE and HKE, c. 2.5% - IndE thus clearly overuses the progressive with stative verbs, while the other two ESLs underuse it. The multiple DCA verifies that IndE diverges from all the other varieties in that six of the twelve lemmata show the strongest deviation from observed to expected in IndE specifically (*existing, having, knowing, residing, staying,* and *understanding*). *Having*, in particular, is characteristic of IndE (see also Balasubramanian, 2009 and Ziegeler & Lenoble, 2020): the progressive usage remains low (less than 1.2%) for all other varieties, whereas the corresponding figure for IndE is 9%. Moreover, most of the instances of BE *having* in IndE extend the semantic boundaries of stative progressives; rather than depicting temporary states, as in (1), IndE regularly (in approximately 75% of all instances) employs BE *having* when referring to non-delimited states, as in (2).

(1) He said India and the United States **are having** good bilateral relations and … (ICE-IND, S2B-004)

(2) The magnet **is having** maximum attraction at its poles. (ICE-IND, S1B-019)

The random forest analysis performed on the whole database indicates that the most important predictor variables are GROUP, LEMMA and FORM (Mtry=2, ntree=1,000, C index 0.92). The ctree, however, shows that LEMMA makes the primary split with {*live, look, reside, stay*} leading to higher probability of the progressive, the behavior of *look* thus overruling the effect of GROUP. The highest probability of the progressive is with *reside* and *stay* in the present or past tense, or with the perfect aspect, in spoken IndE. Looking at ctrees representing individual varieties, LEMMA clearly arises as the most important predictor variable, as it alone makes the only statistically significant split in four of the six varieties, usually splitting *live, look, reside* and/or *stay* from the other lemmata on the path to higher progressive probability. Again, IndE differs from the other varieties in that, next to LEMMA, MODE is involved in a statistically significant split; *exist* and *have*, when occurring in speech, have a slightly higher progressive probability within the 'other' lemmata. Overall, the choice of the progressive over the simple is very much affected by the lexical verb; stance verbs lead to a higher probability of the progressive construction in all varieties.

A brief qualitative look at the data reveals that non-delimited uses of the stative progressive (as in (2) above) do show a statistically significant difference between ENL and ESL varieties (ENL 2% vs. ESL 29%, LL=43.26). We thus need to finetune the original statement with respect to the term “wider use” - it is critical that it is understood in the semantic sense of referring to extension to non-delimited contexts as the sense 'more frequent use of stative progressives in ESLs’ is not confirmed by the present results. Rather, the frequency data, the attraction patterns revealed by the DCA, and the alternation patterns shown by the random forest and ctree analyzes all indicate that IndE differs from the other varieties investigated, including the other ESL varieties, SgE and HKE. It is only within the semantic extension that we find a statistically significant difference between ENLs and ESLs, albeit even then IndE stands out from all other varieties with statistically significant overuse of non-delimited stative progressives (LL=83.83).

### *References*

Aarts, B., Close, J., Leech, G., & Wallis, S. (Eds.). (2013). *The verb phrase in English: Investigating recent language change with corpora*. Cambridge: Cambridge University Press.

Deshors, Sandra C., & Rautionaho, P. (2018). The progressive versus non-progressive alternation: A semantic exploration across World Englishes. *English World-Wide, 39*(3), 309-337.

Hundt, M., & Vogel, K. (2011). Overuse of the progressive in ESL and Learner Englishes - fact or fiction? In J. Mukherjee & M. Hundt (Eds.), *Second-language varieties and learner Englishes: Bridging a paradigm gap* (pp. 145-166). Amsterdam: John Benjamins.

Rautionaho, P., & Fuchs, R. (2020). Recent change in stative progressives: A collostructional investigation of British English in 1994 and 2014. *English Language and Linguistics, 1*, 1-26.

Sharma, D. (2009). Typological diversity in New Englishes. *English World-Wide, 30*, 170-195.

Ziegeler, D., & Lenoble, C. (2020). The stative progressive in Singapore English: A panchronic perspective. In P. Núñez-Pertejo, M. José López-Couso, B. Méndez-Naya & J. Pérez-Guerra (Eds.), *Crossing linguistic boundaries: Systemic, synchronic and diachronic variation in English* (pp. 239-266). London: Bloomsbury Academic.

**Ronan , Patricia**

***TU Dortmund***

**patricia.ronan@tu-dortmund.de**

**Buschfeld , Sarah**

***TU Dortmund***

**sarah.buschfeld@tu-dortmund.de**

**Schneider , Gerold**

***Universität Zürich***

**gschneid@ifi.uzh.ch**

Type of Contribution: Work-in-Progress Report

## Split infinitives in native and non-native speaker data

A corpus-based analysis

*Abstract*

It is well known that language learners do not always acquire complex structures in a native-like fashion. Typical phenomena that are observable in learner language are overgeneralization and creative use of language features (Schneider & Gilquin, 2016). These phenomena are often motivated by L1 transfer or general mechanisms of second language acquisition. Split infinitives are one such phenomenon and often found in English learner language. While they are also found in native speaker English, there they are rare phenomenon. Our qualitative and quantitative corpus-based study enquires how usage patterns and frequencies of split infinitives differ between native speakers and language learners.

Split infinitives are by no means a new introduction into the English language. They can already be found in the Middle English period (Calle-Martin, 2015). However, a strong prescriptivist tradition has existed against them

(Perales-Escudero, 2010), which is now easing, and in contemporary language use, specific collocations are found increasingly, in particular in academic discourse (Perales-Escudero, loc. cit.). An example is

(1) (…) it might be a good thing to actually sort of provoke a question with architects because they do specify these things (…) (BNC, 1994, D 97)

As our source for native speaker data, we use the British National Corpus 1994 (Aston & Burnard 1998). For earlier historical developments the COHA Corpus (Davies, 2010), as well as the Brown family of corpora and the LOB family of corpora, provide insights. Structures and frequencies of split infinitives are compared to learner data taken from the 3.8-million-word International Corpus of Learner English, ICLE (Granger, 2009). To allow genre comparability, the essay-based ICLE data are compared to the related BNC genres of letters, essays and fictional prose, totalling about 18 million words. Data are extracted from the corpora by detecting interceding elements between infinitive marker *to* and the infinitive with the query '\bto_TO [^_]+_RB [^_]+_VB\b'. The queries have been carried out through the web corpus interface Dependency Bank (Lehmann & Schneider 2012).

The results on the native speaker data show that the frequency of split infinitives has been rising gradually during the Late Modern- and Present-Day English periods. The COHA data show low frequencies of split infinitives (0.1/10,000 words) until the mid-19th century, rising slightly in the late 19th century to 0.3/10,000 words, and then declining to 0.1 again until 1940. After this, figures rise to more than 0.9/10,000 words in 2000. The AmE Brown corpus family confirms this trend, showing an increase from 0.27/10,000 in 1930 to 0.75 in 1990. The rise in the BrE LOB corpus is slower, but equally perceivable: from 0.17/10,000 in 1930 to 0.55 in 1990.

In addition to temporal developments, genre differences play a large role in the frequencies of split infinitives. Table 1 shows the distribution of frequencies in select genres in the BNC.

| BNC 1994 | Word total | Mean/10,000 | Standard deviation within the genre |
|---|---:|---:|---:|
| Spoken | 10,641,245 | 1.44 | 0.52 |
| Written | 92,699,909 | 0.4 | 0.28 |
| Letters (personal) | 56,046 | 0.178 | -- |
| Fictional prose | 17,054,264 | 0.232 | -- |
| Essays university | 58,770 | 0.51 | -- |
| Essays school | 154,596 | 1.229 | (> 2 stdv. from written genres) |
| Emails | 223,443 | 1.343 | (> 2 stdv. from written genres) |

Table 1: Frequencies of split infinitives in select genres in the BNC.

Split infinitives are particularly frequent in BNC spoken genres and lower in written genres overall. They are infrequent in personal letters and most frequent in emails and school essays, where their frequencies differ by more than two standard deviations from the mean of the written genres.

Concerning collocational frequencies, the BNC data confirm Perales-Escudero's (2010) observation that split infinitives are particularly frequent in specific collocations. In our BNC data, these are collocations particularly involving the adverb *actually*, especially in *to actually get/do/go/put/say/be*. Also frequent is *just* in *to just go/sit/get*. Further adverbs used with high frequency are *really, even* and *fully.*

Results for learner language on the basis of the ICLE show that learners in the ICLE corpus use significantly more split infinitives, mean= 0.69, than in the BNC across the mean of all written genres (p < 0.0001 according to chi-square), though significantly fewer than in BNC school essays (p = 0.02 according to chi-square with Yates continuity correction). However, the learners' L1 has a profound influence (Table 2).

| L1 | words | frequency /10,000 | Standard deviation among different L1 |
|---|---:|---:|---:|
| All | 3,939,048 | 0.69 | 0.425 |
| Czech | 208634 | 0.096 | n/a |
| French | 214619 | 0.28 | n/a |
| Spanish | 210562 | 0.29 | n/a |
| German | 248666 | 0.8 | n/a |
| Swedish | 290358 | 1.1 | n/a |
| Finnish | 200835 | 1.4 | n/a |
| Norwegian | 221578 | 1.8 | n/a |

Table 2: Use of select split infinitives in ICLE data according to L1.

Czech L1 speakers in the ICLE data use fewest split infinitives, speakers of French and Spanish also use few, L1 speakers of the North Germanic languages Swedish and Finnish use particularly many split infinitives, which may be due to the presence of comparable splittable non-finite verb structures in these languages.

Concerning usage patterns, the ICLE data show innovative use of rare grammatical patterns by learners in the use of copular verbs and many diverse instances of interceding high frequency tokens *not* and *fully*, or multiple interceding elements (example 2), which are both rare in native speaker data.

(2) Motorists cyclists and pedestrians are expected *to always and everywhere* reckon with the sudden appearance of an unguarded child … (ICLE, GEAU3085)

Thus, the learner language in ICLE indeed shows creative and innovative use of the still rare usage pattern of split infinitives, but the influence of the L1 and its own typological patterns seems paramount.

In conclusion we can say that on the basis of the overall comparison of ICLE and BNC data, speakers of British English use few split infinitives in written genres except for emails, which were a new genre at the time, and except for school essays, which may illustrate an increasingly less prescriptive teachers' attitude to split infinitives. Another explanation might be the increasing transatlantic influence on British English, as American English is in the lead in split infinitive use. Average frequencies in ICLE learner language are higher than in written language in the BNC overall, and notably higher than in university essays, but markedly lower than in BNC school essays on average. Speakers of L1s which also offer a possibility to split an infinitive marker from an infinitive, and to place interceding adverbs between these, show high frequencies of split infinitives in English (Swedish, Norwegian). Speakers of L1s in which this is not possible broadly show lower frequencies of split infinitives in our data. However, it may also be the case that a broad orientation of school systems and of language learners towards British or American English, as well as further attitudinal and typological features may play a role here.

## *References*

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Calle-Martin, J. (2015). The split infinitive in Middle English. *NOWELE, 68*(2), 227-250.

Davies, M. (2010-). The Corpus of Historical American English (COHA): 400 million words, 1810-2009. Retrieved from: https://www.english-corpora.org/coha/.

Granger, S. (Ed.). (2009). *International Corpus of Learner English*. Université catholique de Louvain, Centre for English Corpus Linguistics.

Lehmann, H. M., & Schneider, G. (2012). BNC Dependency Bank 1.0. In S. Oksefjell Ebeling, J. Ebeling, & H. Hasselgård (Eds.), *Aspects of corpus linguistics: Compilation, annotation, analysis*. Vol. 12. Helsinki: Varieng.

Perales-Escudero, M. D. (2010). To split or to not split: The split infinitive past and present. *Journal of English Linguistics, 39*(4), 313-334.

Schneider, G. & Gilquin, G. (2016). Detecting innovations in a parsed corpus of learner English. *International Journal of Learner Corpus Research, 2*(2), 177-204.

**Šaldová, Pavlína**

***Charles University***

**pavlina.saldova@ff.cuni.cz**

Type of Contribution: Full Paper

## Patterns of single postpositive adjectives in English

*Abstract*

Adjectives appear after the head noun when they are complemented (heavy AdjP). Single postpositive adjectives (light AdjP) in English are treated, in synchronic grammars, as an exception to the basic syntactic rule that adjectives precede nouns (*a black swan* x **a swan black*) and the postpositive use of a light AdjP is described as subject to severe restrictions (e.g. "adjectives in -*able* or -*ible*, like *suitable* and *possible*, require an attributive superlative or *only*: compare *the best result possible* and **the result possible*" (Huddleston & Pullum 2002, p. 445) or the temporal semantics of the head noun (*years past*)).

Focusing on adjectives which can appear as light AdjPs both in the pre- and post-head positions (*available evidence x evidence available*), the paper aims 1. to identify such adjectives, 2. to examine whether they all follow the same pattern, and 3. to point out the post-head uses which cannot be easily accounted for by the restrictions referred to above.

To retrieve single post-head adjectives from the *BNC* written component (cca 0.25% of all adjective uses in Brown and Frown and cca 8% of all posthead uses (Blöhdorn, 2009)), the query N + AJ0 + V was used. Removing false positives yielded a sample of 1.384 combinations N + Adj, representing 127 adjectives as types, whose frequency displays the Zipfian curve, with ten most frequent adjectives (*concerned*, *available, proper*, *responsible, present*, *payable*, *outstanding*, *possible*, *corporate*, *necessary*) accounting for 84% of all tokens. Sorting the adjectives by word-formation types, -*able*/-*ible*, and (*un*-)-*ed* adjectives dominate.

Sorting all instances into the four groups proposed by Huddleston & Pullum (2002, p. 445), the post-head uses are distributed as follows: *a-*

adjectives (*a young child asleep*) 1%, fixed phrases (*lords temporal*, *pastures new*) 4%, adjectives taking both positions but with a difference in sense (*the people present*, *the city proper*) 47%, and examples of post-head adjectives which may alternate with the pre-head position (*the only day suitable*, *the amount recoverable*) 48%. This broad classification indicates that almost a half of the single postpositive adjective uses are cases where the position is potentially variable.

First, a quantitative approach was adopted to see whether, among the ten most represented adjectives (*corporate* was substituted by *due*) there are differences in their post-head uses in general, i.e. comparing their pre-head uses, and their post-head uses, both light and heavy (*necessary skill*, *do the work necessary*, *the food necessary for his diabetic diet,* respectively). Searching their overall occurrences in the written section of the BNC, the following grouping is observed:

After the head noun, 1. *concerned*, *proper*, *present*, *outstanding*, *possible* are prevailingly used as light AdjPs in 77% (out of 4,304 instances); 2. *responsible, necessary, due, payable* are used, on the contrary, prevailingly (around 82% out of 3,771 instances) in heavy AdjPs; 3. interestingly, *available*, the most frequently used bare adjective without a lexicalized difference overall, shows no clear pattern (n=5,801, 54% heavy vs 46% light AdjPs).

As for the first group (prevalence of light AdjPs), light *concerned* and *proper* contrast with the heavy AdjPs in meaning (*a detective concerned with a recent case*, *clinical societies concerned about brain damage*, *W. continued in a tone proper for instructing young ordinands*) showing a clear difference between the two posthead uses. *Outstanding* is also dominantly used as a light adjective. In the second group (prevalence of heavy AdjP: *responsible, necessary, due* and *payable*) *responsible* instantiates the sense difference in pre- and post-head positions, but the light uses in the post-head position are related to the heavy uses as they can be perceived as elliptical, due to the strong prevalence of the heavy pattern with *for* (1,033 (90%) vs 122). What adjectives appearing dominantly in heavy AdjP share, is that in light uses the complement is easily identifiable in the context, with the adjectival postmodifier contributing to the identification of the referent.

*The most negative aspects of American policing … came together in* <u>*the beating of Rodney King*</u> *in …, which was followed by rioting when* <u>*the white policemen responsible*</u> *were acquitted in April 1992. responsible for the beating*

As was seen with *responsible*, many light post-head uses do not contain explicit constraints identified for *-able*/*-ible* bare adjectives (such as the presence of the superlative or *only*), so a more refined description must be sought.

Two adjectives stand out, *available* and *possible. Available* as the most frequent postpositive adjective without a change in sense is unique in the relatively equal proportion between the post-head positions, contrasting sharply with other adjectives, where one of the post-head patterns dominates and a clear textual link based on a complementation pattern is easy to identify.

Analyzing wider context in instances of *available evidence* and *evidence available*, the reference of the pre-head uses of adjectives is definite and the referent identifiable in the context.

*Kähler rejected any purely historisch* (sic) *approach to the figure of Jesus for three main reasons. First of all,* <u>*the available evidence*</u> *is not of the right kind.* <u>*The gospels*</u> *do not furnish us with the materials for ..*

On the other hand, in the light post-head uses, the referent was not directly present in the context when the light AdjP was used, indicating the reference as non-specific.

Examples with *possible* and other adjectives used without the superlative or *only* contain some lexical expression of restriction/limit and expressions of quantity (e.g. *VHF further* <u>*restricts*</u> *the coverage* <u>*possible, reduce*</u>).

It has been shown that certain light postpositive adjective uses cannot be fully explained by the restrictions indicated above (temporary meaning, presence of superlatives, only, ordinals).

In some cases, the light use represents an ellipsis of the complement (*responsible*), with other adjectives the latent complement cannot be easily identified in previous discourse and the noun phrase may express nonspecific reference. Moreover, in many other examples where the assumed constraints on *-able*/ *-ible* are not present, lexical patterns are identifiable which contain items similar in meaning to *only* (*limit*, *reduce*) and also expressions of quantity (*the amount/rate of*). The question remains as to what extent the patterns identified above overlap and whether they represent one construction.

*References*

Blöhdorn, L. (2009). *Postmodifying attributive adjectives in English: An integrated corpus-based approach*. Frankfurt am Main: Peter Lang.

Bolinger, D. (1952). Linear modification. *PMLA* 67, 1117-1144.

Cinque, G. (2010). *The syntax of adjectives: A comparative study*. Vol. 57. MIT press.

Ferris, C. (1993). *The Meaning of syntax: A study of the adjectives in English*. Harlow: Longman.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

James, D. (1979). Two semantic constraints on the occurrence of adjectives and participles after the noun in English. *Linguistics* 17, 687-705.

Matthews, P. H. (2014). *The positions of adjectives in English*. Oxford: Oxford University Press.

Sanchez-Stockhammer, Christina

*LMU Munich*

**christina.sanchez@lmu.de**

Tochtermann, Johannes

*LMU Munich*

**johannes.tochtermann@campus.lmu.de**

Type of Contribution: Software Demonstration

## WordValue

Simultaneous corpus searches and information mapping on words in context

*Abstract*

### Relevance

Linguistic research commonly results in detailed information that is related to specific words or constructions (e.g. semantics, structure or historical origin). When such item-specific information is mapped onto complete texts, it can reveal the extent to which particular types of text (e.g. specific registers, or texts by individual authors) are characterized by certain types of feature. Analyses of this kind require (1) corpus frequency searches for large numbers of linguistic items, (2) mapping of the text frequencies onto the relevant search items and (3) visualization of the results. If contextual distribution is to be considered, this last point becomes particularly important.

### Contribution of WordValue

Many corpus concordance tools (e.g. AntConc) and corpus query interfaces (e.g. CQPweb) exist, but these are often limited to one search at a time, so that the retrieval of frequency information for large numbers of linguistic items would require labour-intensive consecutive manual searches. Some software packages like #LancsBox and MonoConc permit batch searches for multiple items, but the functionality may be restricted to unlemmatized word forms (see #LancsBox manual: 35) or integrated into highly complex tools (e.g. the MonoConc Manual extends over 149 pages). Our free web application WordValue (www.wordvalue.gwi.uni-muenchen.de) therefore offers a useful complementation to existing software. As an easy-to-use standalone tool with specialized functionality, it is directed at linguistic and non-linguistic users alike. WordValue permits lemmatized and unlemmatized simultaneous corpus frequency searches for predetermined lists of linguistic items. Most importantly, WordValue's graphical interface presents word-related information through colour-coding in the highly informative linguistic co-text, so that the contrastive distribution of search items with similar or identical values can be accessed intuitively at a single glance.

## Features

After setting up a free account at www.wordvalue.gwi.uni-muenchen.de, users can create and store new experiments by uploading two documents with UTF-8 encoding: (1) a corpus text in txt format and (2) an appropriately formatted comma-separated csv spreadsheet with unique search items (= the keys) in the first column and relevant information related to the keys (= the attributes, e.g. language of origin, with their corresponding values like *Germanic* and *Romance*) in the following columns (cf. Table 1). A header is required to identify the keys and the attributes for further processing.

| **keys** | **language_of_origin** | **age** |
|---|---|---|
| game | Germanic | OldEnglish |
| food | Germanic | OldEnglish |
| role | Romance | Later |
| practice | Romance | MiddleEnglish |

Table 1: Extract from an example csv file.

While there are no restrictions on the size of the txt and csv files, the time required for the first run of the experiment increases with both documents' number of lines, as WordValue automatically computes the frequency of all keys from the csv search list (csv) and of all the tokens in the txt corpus for further use. These are then filtered by applying the search options (1) case-sensitivity and (2) lemmatization. The results of the matching process between keys and text tokens can be viewed (1) as a spreadsheet with frequency information or (2) in the innovative rainbow text format.

In rainbow text format, the values of the attributes are mapped directly onto the key items in context by using a customizable colour-coding scheme. The attribute to be colour-coded is selected by means of a drop-down menu. For categorical attributes like *part of speech* (whose values *noun*, *adjective*, *verb* etc. have no inherent ordering), WordValue generates a list of all possible value types. Users can then interact with a colour picker offering a large number of colours (whose precise form depends on the browser used) to assign a colour to each value. Ordinal attributes like *comparison* (whose possible values *positive*, *comparative* and *superlative* are ordered but not equally spaced) can be colour-coded like categorical variables by applying an intuitive systematic colour-coding scheme going from warm to cold colours, from light to dark colours etc. For numerical attributes like *word frequency,* WordValue offers the possibility to map values of these attributes to a colour gradient, for which users can define the start and end colour. WordValue then automatically retrieves all unique values of that attribute and maps these values onto the colour gradient.

For convenient viewing, the table and the rainbow-coloured text on the results page adapt dynamically to any changes in the parameters. The html format permits easy copy-pasting of the coloured texts into Microsoft Word and of the tables into Excel, where they can be saved and further processed by applying e.g. filtering and sorting processes.

### Implementation

Word Value is implemented with a combination of the programming languages Javascript and Python (Version 3.6). It uses spaCy, a natural language processing library, to automatically tokenize and lemmatize given texts. Word Value then loops over all tokens and lemmas from the text and keys from the csv file to find matches. The computational complexity and thus the time required for the computation of results is directly dependent on the number of tokens found in a text, as well as the number of keys present in the csv file. For each combination of token and key, WordValue will check if the token either fully matches a key or if it is part of a sequence of tokens that fully match a key. If a match is found, the respective token is marked as belonging to the matching key. WordValue's search algorithm finds matches even if the search key is separated by spaces, hyphens or line breaks, and it is flexible enough to allow for search keys that are comprised of sequences of keys (e.g. *one of the*).

### Possible applications

Since the mapping process is so flexible, WordValue can be used for very different types of corpus-based research: thus for a word list with part-of-speech information, the software can colour all nouns green, all adjectives red and all verbs blue. This permits a very fast and intuitive identification of combinatorial patterns and their relative distance to each other within the actual context. Another potential field of application besides linguistics and corpus-based literary studies is that of language learning: based on customized vocabulary lists that specify the difficulty level (defined e.g. as progression in a textbook), WordValue can aid the selection of appropriate texts for specific groups of language learners by coding various levels of already mastered vocabulary with different colours. Similar functions exist for preloaded vocabulary lists and corpus-based frequency bands (cf. https://www.lextutor.ca/vp/), but WordValue goes beyond previous software by permitting the use of customized word lists.

### Roadmap: Open points

Potential future avenues to enhance WordValue's functionality include the following options: specification of a colour for all tokens that are not matched by a key in order to maximize contrast; a convenient export option for results; preset csv files for general use; design changes that offer more dynamic and responsive result tables, and automatic part-of-speech tagging of tokens.

### *References*

Brezina, V., Timperley, M., & McEnery, A. (2018). *#LancsBox* v. 4.x. [software package]. *#LancsBox 4.5 manual*.

Barlow, M. (2003). *MonoConc Pro manual*.

Cobb, T. (2020). *Compleat Lexical Tutor* v.8.3. *VocabProfilers*.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

**Schätzle, Christin**

***University of Konstanz***

**christin.schaetzle@uni-konstanz.de**

**Booth, Hannah**

***University of Konstanz***

**hannah.booth@uni-konstanz.de**

Type of Contribution: Full Paper

# Investigating interactional syntactic change in Middle English

## Insights from visual analytics

*Abstract*

The idea that language change results from multiple interacting factors is long-standing in historical linguistics (e.g. Labov, 1963; Malkiel, 1967; Weinreich et al., 1968). Change can be the product of interacting language-internal (i.e. system-driven) and language- external (i.e. socio-political) factors, or result from multiple interacting, exclusively lan- guage-internal factors. *Syntactic* change is taken to interact with changes at other lin- guistic dimensions (e.g. phonology, morphology, semantics) and has thus been labelled an interface phenomenon (Keenan, 1994; Longobardi, 2001). Moreover, interactional change *within* the syntactic system has been addressed from various theoretical per- spectives, e.g. via underlying parametric change in the Principles and Parameters approach (e.g. Kroch, 1989; Lightfoot, 2013), and in the usage-based paradigm under the notion of 'multiple source constructions' (van de Velde et al., 2013).

Alongside increased interest in interactional syntactic change, corpus-based method- ologies specific to the problem have become increasingly sophisticated, yielding many novel findings (e.g. Hilpert & Gries, 2016; Pintzuk et al., 2017). Nevertheless, the tools for investigating syntactic interactions in

diachronic corpus-data remain insuffi- cient. The standard procedure is to extract patterns from corpora via programming scripts or specific query tools, in order to generate data tables containing co-occurrence frequencies and statistical calculations. Finding patterns in such tables is challenging; various feature interactions have to be examined across several tables while taking into account a temporal component. Statistical testing can be useful for quantifying change and validating given hypotheses. However, it is less suited for uncovering syn- tactic change *per se*, since the precise factors involved cannot always be anticipated. Additionally, historical data is naturally sparse, which in turn might render statistical significances delusive.

In this paper, we show how visual analytics (Keim et al., 2008) offers a fruitful ap- proach to investigating interactions between changing syntactic phenomena over time via the HistoBankVis visualization system for historical linguistics (Schätzle et al., 2017; Schätzle et al., 2019). As a showcase, we examine word order change underway during the Middle English (ME) period (c.1100-1500), specifically the loss of the verb-second (V2) constraint and the emergence of S(ubject)-V(erb)-O(bject) word order. Various factors have been claimed relevant for this change (e.g. Los 2009, van Kemenade 2012), but the precise nature of their interactions remains elusive.

Early English texts exhibit a good deal of variation with respect to clausal word order. Old English (OE) is often characterized as a V2 language (e.g. Holmberg, 2015), but there is evidence that V2 was not fully consolidated. A relevant factor is the category of clause-initial constituent (van Kemenade, 1987; Pintzuk, 1999). In clauses with an initial *wh*-phrase, negator or discourse adverb (e.g. *þá* 'then'), henceforth 'Group 1' contexts, the subject is typically postfinite (subject-verb inversion), indicating V2, e.g. (1).

(1) þa cwæþ **he** to him
then said he to them
'then he said unto them...' (BlHom-11:119.49.1511, van Kemenade, 2012)

By contrast, in clauses where some other type of non-subject is clause-initial (e.g. a PP adjunct or object NP), henceforth 'Group 2' contexts, there is a split by subject type. Lexical (i.e. non-pronominal) subjects are typically postfinite, e.g. (2-a), but pronominal subjects are typically prefinite, reflecting an SVO system, e.g. (2-b).

(2) a. [On twam þingum] hæfde **God** þæs mannes sawle gegodod
in two things had God the man's soul endowed
'With two things God had endowed man's soul' (ÆCHom I, 1.20.1, van Kemenade 2012)

b. [Be ðæm] [**we**] magon suiðe swutule oncnawan ðæt...
by that, we may very clearly perceive that
'By that, we may perceive very clearly that ...' (CP 26.181.16, van Kemenade, 2012)

Throughout ME, the subject-verb inversion pattern decreases in frequency, as subjects overall become increasingly prefinite. Various factors have been connected with this. Firstly, the Group 1/Group 2 distinction is said to remain relevant: in Group 1 contexts, subject-verb inversion persists ('residual V2' in Present-day English, Rizzi 1996), while in Group 2 contexts inversion is gradually lost. Specifically, subjects with cer- tain properties (lexical, discourse-new) - which were typically postfinite in OE/early ME - become increasingly prefinite, in line with prenominal and discourse-given subjects (Haeberli, 2002; van Kemenade & Westergaard, 2012). Dialect has also been shown to interact with rates of subject-verb inversion in ME (Kroch & Taylor, 1997; Kroch et al., 2000), specifically that in certain Northern texts both lexical and pronominal subjects invert across the board, indicating a more generalized V2 syntax compared to other regional varieties.

Using data from the Penn-Helsinki Parsed Corpus of Middle English (PPCME2, Kroch & Taylor, 2000), extracted via CorpusSearch queries (Randall, 2005), we examine how subject position (prefinite/postfinite) interacts over time with the various features suggested in the literature:[1]

- subject type: pronominal/lexical

[1] We restrict our investigation to matrix clauses containing a finite verb and an overt subject.

- subject's information-structural status: given/new[1]
- clause-initial constituent:
  - neg/discourse adverb (Group 1)
  - PP/(non-discourse) adverbial/NP object (Group 2)
- dominant dialect of text: north/west-midlands/east-midlands/south[2]

While previous studies look at the interaction between subject position and one or two factors, HistoBankVis allows us to assess interactions between several factors at once. HistoBankVis provides exploratory and flexible access to complex data via three visual- ization components: (i) a compact matrix which allows the monitoring of data sparsity is- sues by providing an overview of the data across time periods, (ii) difference histograms which grant access to differences between data distributions across time, and (iii) di- mension interactions based on the Parallel Sets technique (Bendix et al., 2005; Kosara et al., 2006), for exploring the interrelation between potentially interacting factors.

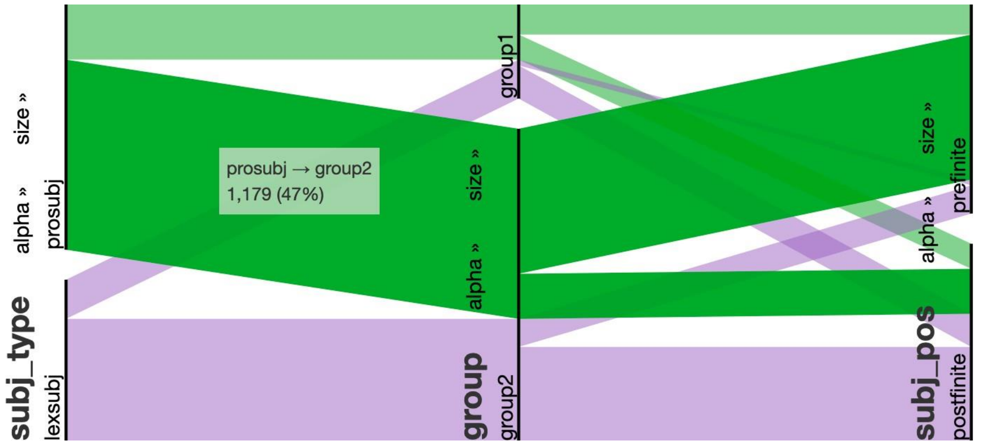


Figure 1: Dimension interaction between subject type (lexsubj/prosubj), clause-initial category (group1/group2) and subject position (prefinite/postfinite) from 1150-1250.

As a first step, we investigate previously suggested interactions via the dimension interaction component (i.e. parallel sets). Firstly, we look at the

[1] The PPCME2 does not annotate for information structure, so we used an approximation: as 'given' we took any subject which is pronominal or has overt definite marking; as 'new' we took any subject which is not pronominal and does not have overt definite marking.

[2] We follow the dialect classification on the corpus website: https://www.ling.upenn.edu/ hist-corpora/PPCME2-RELEASE-4/info/texts-by-dialect.html.

interaction between subject position, subject type and clause-initial category. The insights from HistoBankVis confirm that pronominal subjects lead the change towards becoming increasingly prefinite, with lexical subjects lagging behind. The previously claimed divergence between Group 1 and Group, however, is less clear-cut than often suggested. Compared to Group 2, Group 1 contexts are indeed conservative with respect to the increasing preference for prefinite subjects. Yet the dimension interactions show that Group 1 contexts are far from static: in the earliest period (1150-1250), see Figure 1, more than half of the pronominal subjects are already prefinite. Moreover, we find evidence that lexical subjects in Group 1 also follow suit by at least the third period (1350-1420). We sug- gest that these more nuanced findings for Group 1 are revealed through our decision to exclude *wh*-questions from the group. Furthermore, this change coincides with a striking decrease in clause-initial negation (neg), see the difference histogram in Figure 2. This suggests future investigations should focus on clause-initial discourse adverb (da) contexts for further insights.

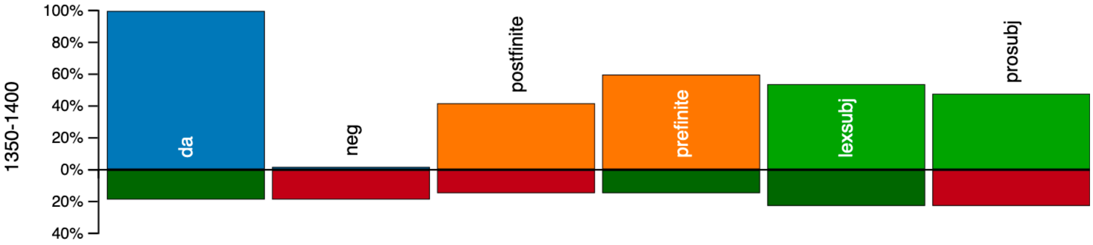


Figure 2: Difference histogram for Group 1 categories (da/neg), subject position and subject type from 1350-1420.

With respect to information structure, the previously suggested correlation between postfinite position and discourse-new subjects, and prefinite position and discourse- given subjects, is borne out overall. Moreover, this correlation weakens over time: by the last period (1420-1500) there is little information-structural effect on position. Since HistoBankVis allows us to investigate multiple factors at once, we can take the investigation further, examining whether information structure is in fact the driving force behind the differences between Group 1 and Group 2. It is difficult to isolate information structure from subject type as a factor, since given subjects will often be pronominal, and new subjects often lexical. Nevertheless, adding both factors into the dimension interaction indicates that subject type is more important than information structure, since the subject type correlation with subject position seems to be stronger overall.

Finally, we add in dialect as a factor, to examine the claim that Northern texts are particularly conservative to the change. Data sparsity is an issue here, with only one Northern text unambiguously dated (*Northern Prose Rule of St. Benet*, 1350-1420). This text has been shown to exhibit a generalized V2 system, with high levels of subject-verb inversion (Kroch and Taylor 1997, Kroch et al. 2000). However, when we include the other Northern texts, which are less clearly dated, the picture is much more variable, with higher rates of prefinite, in particular pronominal, subjects. The quick and easy comparison provided by HistoBankVis thus indicates that the trends across Northern texts from the period are particularly nuanced, and merit further research.

## *References*

Bendix, F., Kosara, R., & Hauser, H. (2005). Parallel sets: Visual analysis of categorical data. In *IEEE Symposium on Information Visualization INFOVIS*, pp. 133-140.

Haeberli, E. (2002). Observations on the loss of verb-second in the history of English. In C. Zwart & W. Abraham (Eds.), *Studies in comparative Germanic syntax* (p. 245). Vol. 53. Amsterdam: John Benjamins.

Hilpert, M., & Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. In M. Kytö & P. Pahta (Eds.), *The Cambridge handbook of English historical linguistics* (pp. 36-53). Cambridge: Cambridge University Press.

Holmberg, A. (2015). Verb second. In T. Kiss & A. Alexiadou (Eds.), *Syntax. Theory and analysis: An international handbook* (pp. 342-383). Berlin: Mouton de Gruyter.

Keenan, E. (1994). Creating anaphors: An historical study of the English reflexive pronouns. Ms. UCLA.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In A. Kerren, J.T. Stasko,

J. D. Fekete, & C. North (Eds.), *Information visualization* (pp. 154-175). Berlin: Springer.

Kosara, R., Bendix, F., & Hauser, H. (2006). Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, *12*(4), 558-568.

Kroch, A. & Taylor, A. (2000). The Penn-Helsinki parsed corpus of Middle English (PPCME2). 2nd edition. Department of Linguistics, University of Pennsylvania.

Kroch, A., Taylor, A., & Ringe, D. (2000). The Middle English verb-second constraint: A case study in language contact and language change. In S. C. Herring, P. van Reenen & L. Schøsler (Eds.), *Textual parameters in older language* (pp. 353-391). Amsterdam: John Benjamins.

Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, *1*(3), 199-244.

Kroch, A. S. & Taylor, A. (1997). Verb movement in Old and Middle English: Dialect variation and language contact. In A. van Kemenade & N. Vincent (Eds.), *Parameters of morphosyntactic change* (pp. 297-325). Cambridge: Cambridge University Press.

Labov, W. (1963). The social motivation of a sound change. *Word*, *19*(3), 273-309.

Lightfoot, D. W. (2013). Types of explanation in history. *Language*, *89*(4), e18-e38.

Longobardi, G. (2001). Formal syntax, diachronic minimalism, and etymology: The history of French *chez*. *Linguistic Inquiry*, *32*(2), 275-302.

Los, B. (2009). The consequences of the loss of verb-second in English: Information structure and syntax in interaction. *English Language and Linguistics*, *13*(1), 97-125.

Malkiel, Y. (1967). Multiple versus simple causation in linguistic change. In *To honor Roman Jakobson: Essays on the occasion of his seventieth birthday* (pp. 1228- 1246). Hague: Mouton de Gruyter.

Pintzuk, S. (1999). *Phrase structures in competition: Variation and change in Old English word order*. New York: Garland.

Pintzuk, S., Taylor, A., & Warner, A. (2017). Corpora and quantitative methods. In A. Ledgeway & I. Roberts (Eds.), *The Cambridge handbook of historical syntax* (pp. 218-240). Cambridge: Cambridge University Press.

Randall, B. (2005). CorpusSearch2 User's Guide. Philadelphia: Dept. of Linguistics, University of Pennsylvania. Retrieved from: http://corpussearch.sourceforge.net.

Rizzi, L. (1996). Residual verb second and the *wh*-criterion. In A. Belletti & L. Rizzi (Eds.), *Parameters and functional heads: Essays in comparative syntax* (pp. 63-90). Oxford: Oxford University Press.

Schätzle, C., Dennig, F. L., Blumenschein, M., Keim, D. A., & Butt, M. (2019). Visualizing linguistic change as dimension interactions. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (pp. 272-278) Florence: Association for Computational Linguistics.

Schätzle, C., Hund, M., Dennig, F. L., Butt, M., & Keim, D. A. (2017). HistoBankVis: Detecting language change via data visualization. In G. Bouma & Y. Asedam (Eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Lan- guage* (pp. 32-39). Linköping: Linköping University Electronic Press.

van de Velde, F., De Smet, H., & Ghesquière, L. (2013). On multiple source constructions in language change. *Studies in Language*, *37*(3), 473-489.

van Kemenade, A. (1987). *Syntactic case and morphological case in the history of English*. Dordrecht: Foris.

van Kemenade, A. (2012). Rethinking the loss of verb second. In T. Nevalainen & E. C. Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 823-834). Oxford: Oxford University Press.

van Kemenade, A. & Westergaard, M. (2012). Verb-second variation in Middle English. In A. Meurman-Solin, M. J. Lopez-Couso & B. Los (Eds.), *Information structure and syntactic change in the history of English* (pp. 87-118). Oxford: Oxford University Press.

Weinreich, U., Labov, W., & Herzog, M. I. (1968). Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics* (pp. 95-195). Austin: University of Texas Press.

**Schneider, Gerold**

***University of Zurich***

**gschneid@ifi.uzh.ch**

Type of Contribution: Full Paper

# Do non-native speakers read differently?

## Correlations between language models and reading times of native and non-native eye-tracking data

*Abstract*

**Introduction**

In cognitive linguistics, the influence of quantitative data and models is paramount (Janda 2013), frequency and frequency-derived measures show strong correlations to processing time and are also seen as primary factors in language change (Bybee 2007). Entrenched, formulaic sequences are easier to process for native speakers (Conklin and Schnitt 2012), but more difficult to learn for L2 learners (Schneider and Gilquin 2016). Formulaicity is also related to Sinclair's idiom principle (Sinclair 1991), which can be measured by surprisal (Levy and Jaeger 2007, Schneider and Grigonyte 2018). We investigate the correlation between reading times as manifested in eye- tracking corpora and text-derived measures of formulaicity, e.g. surprisal, word frequency (Smith and Levy 2013), how well models can predict then, and what the differences between L1 and L2 readers are.

**Data and Method**

We use eye-tracking reading times (RT) from a) Reading Times For Model Evaluation (Frank, Frank et al. 2013) and b) the Ghent Eyetracking Corpus (GECO, Cop et al. 2017). The latter has a substantial portion of L2 readers. We compare 4 L1 readers from Frank, 12 L1 readers and 7 L2 readers from GECO.

**Results**

Individual variation is very strong, and reveals that fast readers have a higher correlation to surprisal. L1 readers are faster, and correlate better to surprisal than L2 readers, as Figure 1 shows. In order to obtain smoother data, we mainly correlate and predict each word's RT summed over all readers in the same group (L1 or L2). The correlation of this RT mean to surprisal is 0.35 for L1, and 0.25 for L2. The only better correlate we found is word length, with a correlation of 0.58 for L1, and 0.47 for L2.

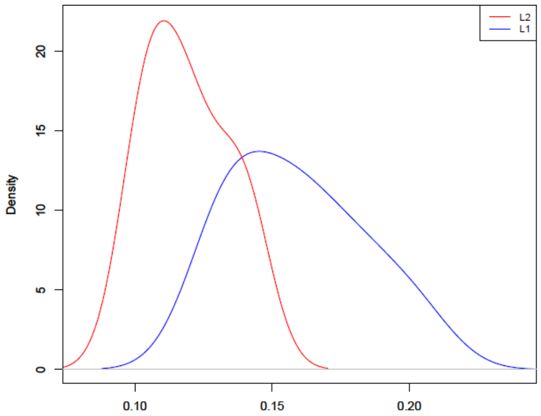


Figure 1: Correlations between reading time of L1 and L2 individuals and surprisal, from GECO.

Predicting RT mean with linear regression reveals the following hierarchy of factor weights:

**Word Length > Punctuation ≈ Surprisal > POS Tag**

The order factor weights stays identical for the prediction of L1 and L2 RT, as Figure 2 illustrates.

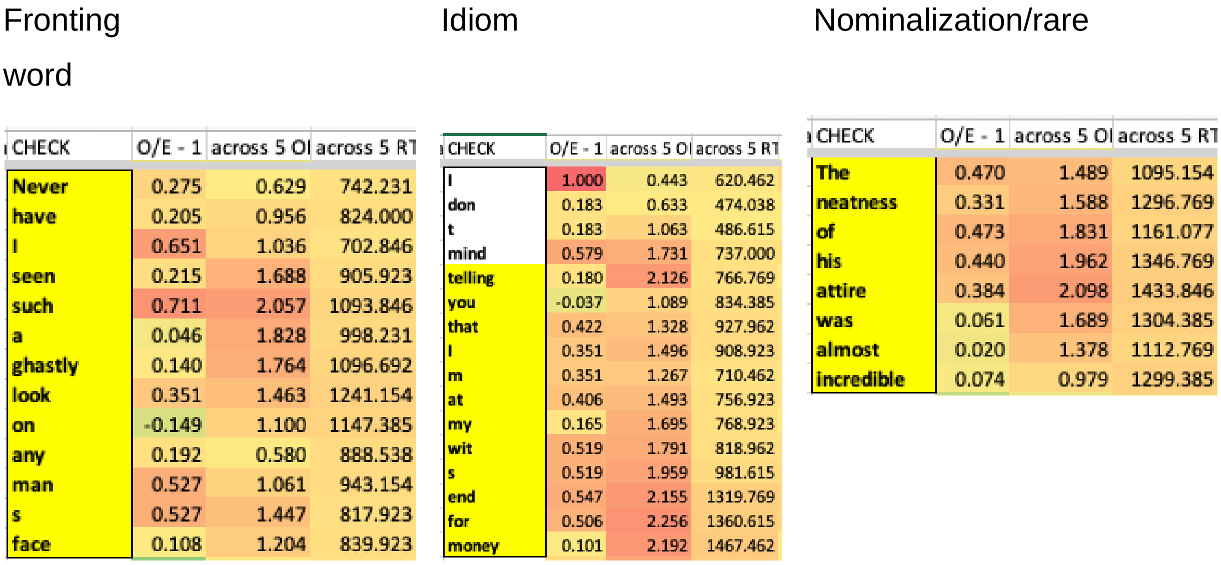
```
> summary(gecofit1s) ## L1
Df        Sum Sq          Mean Sq F value Pr(>F)
LENGTH          1        107561422  107561422  6678.1    <2e-16  ***
SURPRISAL       1        3377679    3377679    209.7     <2e-16  ***
PUNCTUATIO      1        4674232    4674232    290.2     <2e-16  ***
N
Residuals       12745    205279403  16107
```

```
> summary(gecofit2s) ## L2
                Df       Sum Sq     Mean Sq    F value   Pr(>F)
LENGTH          1        138631778  138631778  3663.56   < 2e-16     ***
SURPRISAL       1        1679436    1679436    44.38     2.81e-11    ***
PUNCTUATIO      1        2818953    2818953    74.50     < 2e-16     ***
N
Residuals       12745    482280309  37841
```

Figure 2: Linear regression to predict RT mean per word for L1 and L2.

Predictions of RT means by the linear models are off by typically 40% for L1 and 45% for L2 words. Including L1 RT to predict L2 RT shows that word length is a better predictor than L1 RT.

Finally, we zoom in on areas where L2 readers spend considerably more time to read. We observe that the following phenomena are often linked to increased L2 RT: fronting, zero-relatives, rare words, nominal style, long attachments, rare constructions and idioms. Examples are given in Figure 3.



Fronting word

| CHECK | O/E - 1 | across 5 OI | across 5 RT |
|---|---|---|---|
| Never | 0.275 | 0.629 | 742.231 |
| have | 0.205 | 0.956 | 824.000 |
| I | 0.651 | 1.036 | 702.846 |
| seen | 0.215 | 1.688 | 905.923 |
| such | 0.711 | 2.057 | 1093.846 |
| a | 0.046 | 1.828 | 998.231 |
| ghastly | 0.140 | 1.764 | 1096.692 |
| look | 0.351 | 1.463 | 1241.154 |
| on | -0.149 | 1.100 | 1147.385 |
| any | 0.192 | 0.580 | 888.538 |
| man | 0.527 | 1.061 | 943.154 |
| s | 0.527 | 1.447 | 817.923 |
| face | 0.108 | 1.204 | 839.923 |

Idiom

| CHECK | O/E - 1 | across 5 OI | across 5 RT |
|---|---|---|---|
| I | 1.000 | 0.443 | 620.462 |
| don | 0.183 | 0.633 | 474.038 |
| t | 0.183 | 1.063 | 486.615 |
| mind | 0.579 | 1.731 | 737.000 |
| telling | 0.180 | 2.126 | 766.769 |
| you | -0.037 | 1.089 | 834.385 |
| that | 0.422 | 1.328 | 927.962 |
| I | 0.351 | 1.496 | 908.923 |
| m | 0.351 | 1.267 | 710.462 |
| at | 0.406 | 1.493 | 756.923 |
| my | 0.165 | 1.695 | 768.923 |
| wit | 0.519 | 1.791 | 818.962 |
| s | 0.519 | 1.959 | 981.615 |
| end | 0.547 | 2.155 | 1319.769 |
| for | 0.506 | 2.256 | 1360.615 |
| money | 0.101 | 2.192 | 1467.462 |

Nominalization/rare

| CHECK | O/E - 1 | across 5 OI | across 5 RT |
|---|---|---|---|
| The | 0.470 | 1.489 | 1095.154 |
| neatness | 0.331 | 1.588 | 1296.769 |
| of | 0.473 | 1.831 | 1161.077 |
| his | 0.440 | 1.962 | 1346.769 |
| attire | 0.384 | 2.098 | 1433.846 |
| was | 0.061 | 1.689 | 1304.385 |
| almost | 0.020 | 1.378 | 1112.769 |
| incredible | 0.074 | 0.979 | 1299.385 |

Figure 3: Examples of fronting, idioms, nominalizations and rare words.

The comparison of parsimonious language models to processing time has a variety of applications in applied linguistics, ranging from cognition, stylistics, automatic style checking and essay grading, to learner language research and language simplification.

## *References*

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Conklin, K., & Norbert, S. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45-61.

Cop, U., Nicolas, D., Drieghe, D., & Wouter, D. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods, 49*(2), 602-615.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193-210.

Frank, S. .L., Monsalve, I.F., Thompson, R.L., & Vigliocco, G. (2013). Reading-time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods, 45*, 1182-1190.

Janda, L. A. (2013). *Cognitive linguistics: The quantitative turn*. Berlin: Mouton de Gruyter.

Levy, R., & Jaeger, F. T. (2007). Speakers optimize information density through syntactic reduction. *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

Schneider, G., & Gaëtanelle, G. (2016). Detecting innovations in a parsed corpus of learner English. *International Journal of Learner Corpus Research, 2*(2), 177-204.

Schneider, G.,  & Gintare, G. (2018). From lexical bundles to surprisal and language models: Measuring the idiom principle on native and learner language. *Applications of Pattern-Driven Methods in Corpus Linguistics, 82*, 15-55.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Smith, N., & Roger, L. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302-319.

**Schützler, Ole**

***University of Bamberg***

**ole.schuetzler@uni-bamberg.de**

Type of Contribution: Full Paper

## The grammaticalization of concessive subordinators in Scottish Standard English

*Abstract*

This paper focuses on the concessive subordinators *although*, *though* and *even though*, comparing their functions in Scottish Standard English (SSE) and two other major standard dialects in the British Isles, Southern British Standard English (SBSE) and Irish Standard English (IrSE).

Apart from noting a less formal character of *though* and the emphatic character of *even though,* standard grammars treat the three connectives as fully equivalent (e.g. Quirk et al., 1985). However, functional differences between them exist and can be captured using mainly two of Sweetser's (1990; cf. Hilpert. 2013) semantic categories, here re-labelled as 'anticausal' and 'dialogic'. In the anticausal concessive of (1), meeting the locals would normally be unlikely due to an implicit conditionality ('**if** no time **then** no meeting'). In the dialogic example (2), one proposition qualifies the other with no obvious causal or conditional relationship between them (also see Schützler, 2020).

(1) Though I had little time to spare I met a number of Shetlanders, the majority of whom were dyed-in-the-wool 'Better Together' supporters. (ICE-SCO editorials-26)

(2) Monday 7th would suit me best for the film, though I could do Thursday 10th if you would both prefer. (ICE-SCO social letters-54)

While all three conjunctions can combine with these semantic types, previous research by the author has shown that, in addition to a general frequency

ranking (*although* > *though* > *even though*), there is a strong link between the use of *even though* and anticausal meaning in many varieties of English.

There is virtually no previous research concerning the functionalities and relative frequencies of concessive conjunctions in Scottish Englishes. For vernacular varieties, bordering on the Scots end of the sociolinguistic continuum, Miller (2008, pp. 318-319) claims that there are no adverbial clauses with *although*, and that coordinated constructions (with *but*) or clause-final *though* are used instead. Häcker (1999, pp. 139-142) states that *though* is more frequent in Scots than *although*, which has retained its emphatic meaning and is thus functionally closer to *even though*. Since I assume that there is at least some parallel between patterns in Scots and standard usage in Scotland, my expectation is that, in the alternation between concessive subordinators, SSE will more often opt for *though* compared to other varieties. This paper not only puts this hypothesis to the test but also investigates whether or not the functional division of labour between conjunctions - with a correlation of intra-constructional semantics and the selection of a marker - is similar in different British-Isles Englishes.

Based on *n* = 17 written genres from the Scottish, Southern British and Irish components of the International Corpus of English (ICE; cf. Kirk & Nelson, 2018), all occurrences of *although*, *though* and *even though* were extracted with AntConc (Anthony, 2018) and annotated for the two semantic types outlined above. For SSE, I draw upon the newly compiled (but still incomplete) ICE-Scotland (Schützler, Gut & Fuchs, 2017). There was a total of *n* = 637 observations in *n* = 303 texts. Data were coded for the outcome **MARKER** (*although*, *though*, *even though*), as well as the predictors **VARIETY** (SSE, SBSE, IrSE), and **SEMANTICS** (anticausal, dialogic). Using the R-package brms (Bürkner, 2020), a Bayesian multinomial mixed-effects regression model with random intercepts for **TEXT** and **GENRE** was fitted to predict the choice of construction. The model was specified thus:

marker ~ variety * semantics + (1 | text) + (1 | genre)

Figure 1 shows that, for dialogic concessives, there is the same general frequency ranking in all three varieties. There are no substantial differences between them, with the exception of IrSE, which uses *even though* much more frequently, if still at a relatively low level. Anticausal concessives indeed increase the probabilitiy of using *even though*, at the expense of the other two conjunctions.

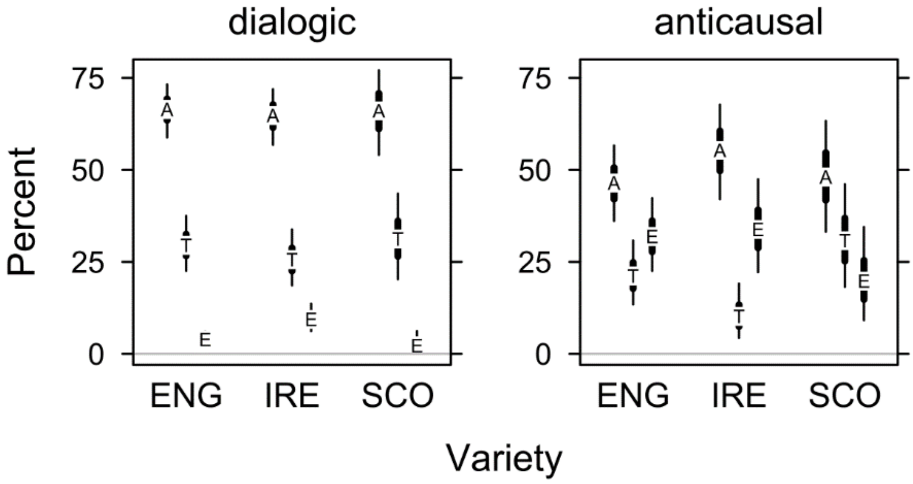


Figure 1: Percentages of markers by variety and semantic type (50% and 90% uncertainty intervals; A = although, T = though, E = even though).

In addition to the patterns described above, SSE stands out in using *though* more frequently and *even though* less frequently than the other two varieties. While SSE, IrSE and SBSE thus share a basic link between semantic factors and the choice of conjunction, SSE draws more strongly on what can be viewed as the etymological origin of all three connectives (*though*), while the most recent addition to the set (*even though*) is given a lesser role. This can tentatively be interpreted as a greater conservatism and a lower degree of grammaticalization of these markers in SSE.

The present research needs to be expanded by (i) including more data from ICE-Scotland and by (ii) inspecting spoken language, too. Furthermore, historical data could shed light on the diachronic development of the correlation between form and function, particularly on the question of whether or not the dialogic type has indeed developed out of the anticausal one (cf. Sweetser, 1990) and whether the relatively uniform selection patterns seen in the left-hand panel of Figure 1 have emerged in this process. This kind of investigation would be possible in SBSE as well as earlier varieties of Scots; for IrSE and SSE, however, we lack the necessary corpus data.

## *References*

Anthony, L. (2018). AntConc. *A freeware corpus analysis toolkit for concordancing and text analysis*. Version 3.5.7 [computer program] Retrieved from: http://www.laurenceanthony.net/software.html

Bürkner, P. (2020). *brms: Bayesian Regression Models using 'Stan'*. R-package version 2.12.0. Retrieved from: https://cran.r-project.org/web/packages/brms/brms.pdf

Häcker, M. (1999). *Adverbial clauses in Scots: A semantic-syntactic study*. Berlin: Mouton de Gruyter.

Hilpert, M. (2013). *Constructional change in English: Developments in allomorphy, word formation, and syntax*. Cambridge: Cambridge University Press.

Kirk, J., & Nelson, G. (2018). The International Corpus of English project: A progress report. *World Englishes, 34*(4), 697-716.

Miller, J. E. (2008). Scottish English: Morphology and syntax. In B. Kortmann & C. Upton (Eds.), *Varieties of English* (pp. 299-327). Vol. 1: *The British Isles*. Berlin: Mouton de Gruyter.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Arnold.

Schützler, O. (2020). *Although*-constructions in varieties of English. *World Englishes* [Early view].

Schützler, O., Gut, U., & Fuchs, R. (2017). New perspectives on Scottish Standard English: Introducing the Scottish component of the International Corpus of English. In J. C. Beal & S. Hancil (Eds.), *Perspectives on Northern Englishes* (273-302). Berlin: Mouton de Gruyter.

Sweetser, E. E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

## Schweinberger, Martin

**The University of Queensland, Australia**

**m.schweinberger@uq.edu.au**

Type of Contribution: Full Paper

# Best practices in Corpus Linguistics

What lessons should we take from the Replication Crisis and how can we guarantee high quality in our research?

*Abstract*

This talk represents a discussion note rather than a research presentation as it describes current and ongoing issues relating to Best Practices (BP) in Corpus Linguistics (CL). The presentation highlights problematic practices currently followed researchers engaged in CL, and proposes workable solutions which help guarantee transparency, replicability, and high quality of research outputs which will also serve to regain public trust. As such, the aim of this talk is to raise awareness about issues relating to BP in CL and to start a discussion of how to guarantee high quality of CL research.

While a discussion about BP, i.e. a method or procedure that is superior to alternatives because it produces results that are more reliable, transparent, replicable, or more ethical, in CL has recently begun (e.g. Berez-Kroeker et al. 2018; Ruhi et al. 2014), this discussion focuses almost exclusively on data compilation whereas issues relating to BP in data processing and analysis remain underexplored.

This paper argues that substantially more attention has to be placed on the repercussions of the Replication Crisis for CL and explores inferences that CL can draw from the Replication Crisis. The Replication Crisis is an ongoing methodological crisis primarily affecting parts of the social and life sciences beginning in the early 2010s (Diener & Biswas-Diener 2019) which contributes to the public loss of trust that Humanities, Social Science, and Arts has experienced of the past two decades (McRae 2018, Yong 2018).

Current problems in CL that negatively affect the transparency of CL research are that analyses are often not reproducible and that the replication of existing research does not take place in a meaningful or substantive frequency.

It is argued that the reasons for these issues are that

- Replications is disincentivized
    - While journals would likely lose readership if similar studies were published repeatedly while researchers enact face-saving strategies as the publication of studies showing conflicting findings could negatively impact career trajectories
- Lack of resources
    - Even if researchers wanted to implement BP, support infrastructure including training and materials around Best Practices are not available to researchers to date
- For individual researchers as well as teams, proposed solutions to these issues include:
- Following FAIR data requirements
    - Data should be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016).
- Publication of data
- clear examples for how to cite data should be provided by corpus compilers
- data should be published in online repositories
- Digital Object Identifiers (DOI) should be assigned to data before studies that are based on the data can be published
- Notebooks
    - R and Jupyter Notebook and their publication on publicly accessible repositories such as GitHib guarantee full transparency and replicability
- Scripts over Tools
    - The use of R or Python should be incentivized as scripts enable replication while the use of software applications (tools such as Sketch Engine, MonoConc, SPSS, etc.) which are often

    - untransparent digital black boxes disincentivize replication to due restricted access and time restraints
- Documentation
    - Documentation of workflows, where to find what, and whom to ask for help are essential not only for on-boarding of new staff but also allow the evaluation of ongoing research practices
- Archiving
    - Completed projects (data, code, etc.) should be stored on publicly available repositories (e.g. GitHub) to allow public access, transparency, and replicability.

Given the highlighted issues, I would like to suggest that as a community, we should stress the importance of replicability and transparency of research by actively promoting replication in teaching contexts. In addition, journal editors as well as reviewers could ask for citation of data sources as well as requesting the submission of data and scripts when reviewing papers. Furthermore, more substantive investments in training and a support infrastructure for issues relating to data management and transparent data analysis options (R, Python, Git, Markdown, etc.) would be required to remedy existing problems. Finally, it is necessary to continue the discussion around BP in CL as well as to form research networks and facilitate panels that help highlighting issues related to BP in CL.

### *References*

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., & Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics, 56*(1), 1-18.

Diener, E.., & Biswas-Diener, R. (2019). The replication crisis in psychology. *NOBA Project.* Retrieved from: https://nobaproject.com/modules/the-replication-crisis-in-psychology.

McRae, M. (2018). Science's 'replication crisis' has reached even the most respectable journals, report shows. *ScienceAlert.* Retrieved from: https://www.sciencealert.com/replication-results-reproducibility-crisis-science-nature-journals.

Ruhi, Ş., Haugh, M., & Schmidt, T. (2014). *Best practices for spoken corpora in linguistic research*. Cambridge: Cambridge Scholars Publishing.

Yong, E. (2018). Psychology's replication crisis is running out of excuses. Another big project has found that only half of studies can be repeated. And this time, the usual explanations fall flat. *The Atlantic.* Retrieved from: https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/.

**Seitanidi, Eleni**

***Lund University***

**eleni.seitanidi@englund.lu.se**

**Simaki, Vasiliki**

***Lund University***

**vasiliki.simaki@englund.lu.se**

**Paradis, Carita**

***Lund University***

**carita.paradis@englund.lu.se**

Type of Contribution: Full Paper

# ‘*We Simply Organic! #WeLoveorganic #organicfood*’

Data compilation and contrastive analysis of consumer texts about organic food

*Abstract*

This study explores the contexts in which English and Greek consumers use *organic* and its Greek equivalent *βιολογικό* when discussing organic food in tweet utterances and hashtags to better understand their range of use and possible cultural differences. Our research is motivated by the lack of consensus regarding the meaning application of *organic* (Haedicke, 2016), which has been found to create confusion among consumers (Anisimova et al., 2019; Fenko et al., 2016). The study first describes the methodology and the data processing, followed by an account of the contextual use of *organic* and *βιολογικό* in the connected text. Next, it briefly explicates the hashtag-related literature and analyzes the uses and functions of *organic* and *βιολογικό* in hashtags in the two data sets.

For the purposes of our study, we have compiled a Twitter corpus, comprising 147,689 running words constituting 10,614 posts in Greek, and 1,653,224 running words found in 118,023 posts in English. The data sets form part of the Sustainable Product Consumer Review (SPCR) corpus compiled for the LangTool project[1] at Lund University, from which we extracted all posts related to organic food and converted the textual data into txt. To ascertain a high reliability level, we removed URLs leading to multiple identical examples running a Python script and maintained all unique examples. Next, we manually removed the remaining non-English and non-Greek posts from the .txt files based on AntConc[2] wordlists.

For each data set, we created a word list of the ten most frequent content words (lemmas) in order to explore in what environments or about what topics they were most often used*.* The scope of the chunk of words was three words before and three words after the concordance words, i.e., *organic* and *βιολογικό.* The results are shown in Table 1.

[1] https://projekt.ht.lu.se/langtool/.

[2] https://www.laurenceanthony.net/software/antconc/.

| English Corpus | | | Greek Corpus | | |
|---|---|---|---|---|---|
| **Lemma** | **Total count** | **Hashtag count** | **Lemma** | **Total count** | **Hashtag count** |
| organic | 102150 | 89954 | βιολογικό (organic) | 11580 | 125 |
| food | 92859 | 82263 | προϊον (product) | 6040 | 9 |
| healthy | 21967 | 19505 | ελαιόλαδο (olive oil) | 540 | 9 |
| vegan | 21721 | 17607 | τρόφιμα (foods/ groceries) | 543 | 2 |
| health | 17104 | 15861 | γάλα (milk) | 527 | 8 |
| vegetarian | 13079 | 12837 | διατροφή | 500 | 13 |
| glutenfree | 12339 | 12328 | μέλι (honey) | 574 | 9 |
| raw | 11424 | 10773 | δέρμα (skin) | 438 | 1 |
| gmo | 9864 | 7721 | λάδι (oil) | 400 | 4 |
| recipe | 9845 | 8951 | αγορά (market) | 719 | 1 |

Table 1: The most frequent lemmas of the English and Greek sets in continuous text and in hashtags.

As indicated by the lemmas in Table 1, the English Twitter posts highlight different dietary styles, while the Greek Twitter posts emphasize the commercial aspect of organic food as indicated by the existence of three items related to the domain of doing business. The commercial nature of the Greek corpus is also reflected through the occurrence of brand names and offers, e.g. *5 λίτρα βιολογικό ελαιόλαδο 23 € ΜΟΝΟ!!!* (‘5 litres of organic olive oil for ONLY 23€!!!’). This commercial focus of the Greek posts may indicate that Twitter is primarily used by businesses in Greece.

Despite this difference in terms of domain focus, the two sets also exhibit similarities. Both corpora contain statements expressing uncertainty about organic food meaning and characteristics, e.g. *What is organic food?* and *Τι είναι τα βιολογικά προϊόντα?* (‘What are organic products?’), which echoes the observation about consumer confusion regarding the meaning of organic food (Anisimova et al., 2019; Fenko et al., 2016). The two sets also contain many comparisons between organic and conventional food, e.g. *#Organic #food is "not healthier"*, and *Βιολογικά Vs συμβατικά τρόφιμα. Οδηγός για έξυπνους* ('Organic vs conventional foods. A guide for the clever'), which supports the observations made by Suciu et al. (2019).

Similarly, in both corpora, post writers express doubts regarding organic food trustworthiness, e.g. *Is your organic food a fraud?* and *Πόσο βιολογικά είναι τα βιολογικά προϊόντα?* ('How organic are organic products?')*,* which is an interesting finding as trust has been highlighted as one of the main factors determining organic food consumption (Anisimova et al., 2019; Teng & Wang, 2015). Having analyzed how speakers refer to *organic* and *βιολογικό* in utterances, we next discuss hashtags, which also contribute to meaning (Laucuka, 2018), albeit in a different formal style.

According to Zappavigna (2015), hashtags have assumed functions beyond their initial topic tracking use. They now also perform experiential, interpersonal and textual functions. Laucuka (2018) proposes ten more specific communicative hashtag functions, i.e. "topic-marking, aggregation, socializing, excuse, irony, providing metadata, expressing attitudes, initiating movements, propaganda, and brand marketing" (p. 56).

As Table 1 shows, Greek users make much less use of hashtags, i.e. 2,301 hits in total, as opposed to 639,795 hits in the English set. This may indicate that hashtag use is more strongly connected to Twitter posting in the English cultural setting than the Greek one. The lemmas found in the English frequency list are very often found in hashtags, while the respective Greek lemmas rarely occur in hashtags.

Table 2 lists some examples of the most common functions of English and Greek hashtags. They include the functions already identified by Laucuka (2018), but we also found two additional functions in the English set, i.e. providing a solution, and expressing contrast. Furthermore, the Greek set contains various branding function examples, which reinforces the observation about its commercial focus mentioned previously.

| Function | Example |
|---|---|
| Topic marking: | *#Spain is the second largest producer of #organic #food in the EU with 27,877 farms* |
| Metadata: | *#JacketPotato #SweetChilli #Chicken #Tomato & #Lettuce #Salad #Organic #Lunch #Dinner #Healthy #EatClean #Food pic.* |
| Expressing attitudes: | *Eat #organic! #greenearthorganics #loveorganic #food #foodie #foodgasm #foodporn #instafood…* |
| Initiating movements - Propaganda: | *Best plant milk ever! #vegan #vegetarian #glutenfree #food #GoVegan #organic #healthy #RAW #recipe #health #whatveganseatpic.*<br>*#FuckMonsanto #GMO #Monsanto #KillingFood #Organic #KillingHumans #killers #serialKillers #food…* |
| Branding - Self-branding: | *Οταν τό μυαλό μας πηγαίνει στά #Βιολογικά #Προϊόντα Μενοίκιο #ΕλληνικάΠροϊόντα #GreekQualityProducts*<br>‘When we think of #Organic #Menoikio Products #GreekProducts #GreekQualityProducts’<br>*#dubai #food #foodie #blogger #organic* |
| Humour - Irony: | *Βέβαια με τα μέτρα οχι βιολογικά ούτε συμβατικά λαχανικά δεν θα απολαμβάνουμε #4ο_μνημονιο #MasterChefGR*<br>‘Of course, with the measures we will not even be able to enjoy conventional vegetables, let alone organic #4th_memorandum #MasterChefGR’ |
| Providing a solution: | *#Regeneration #climatechange #globalwarming #organic #food* |
| Expressing contrast: | *#GMO #Pesticide #Herbicide #Disease #Cancer #Carcinogen #Organic #Food #Fresh #Produce #Health* |

Table 2: Hashtag functions in English and Greek.

To conclude, this study examined how Greek and English Twitter writers use *organic* and *βιολογικό* in their utterances and hashtags with a view to discovering their meaning applications and potential cultural differences. First, the utterance-related analysis reveals a substantial quantitative difference, where English tweets about organic food are much more common than Greek ones. Second, in the English data set, there is a focus on dietary styles, while there is a commercial focus in the Greek data. The tweet hashtags cover quite a wide range of functions in both languages. Most of the functions have been identified in the literature before, but we also found two new uses, namely providing a solution and expressing a contrast.

## *References*

Anisimova, T., Felix, M., & Weiss, J. (2019). Controlled and uncontrolled communication stimuli and organic food purchases: The mediating role of perceived communication clarity, perceived health benefits, and trust. *Journal of Marketing Communications*, *25*(2), 180-203.

Fenko, A., Leone, K., & Bialkova, S. (2016). Overcoming consumer scepticism toward food labels: The role of multisensory experience. *Food Quality and Preference*, *48*, 81-92.

Haedicke, M. A. (2016). *Organizing organic: Conflict and compromise in an emerging market*. Stanford University Press, eBook Academic Collection (EBSCOhost).

Laucuka, A. (2018). Communicative functions of hashtags. *Economics and Culture*, *15*(1), 56-62.

Suciu, N., Federico, F., & Trevisan, M. (2019). Organic and conventional food: Comparison and future research. *Trends in Food Science & Technology*, *84*, 49-51.

Teng, C., & Wang, Y. (2015). Decisional factors driving organic food consumption: Generation of consumer purchase intentions. *British Food Journal*, *117*(3), 1066-1081.

Zappavigna, M. (2015). Searchable talk: The linguistic functions of hashtags. *Social Semiotics*, *25*(3), 274-291.

## Sundberg, Daniel

***Linnaeus University***

**daniel.o.sundberg@lnu.se**

## Tsiviltidou, Zoi

***University of Leicester***

**tsivizoi@gmail.com**

Type of Contribution: Poster Presentation

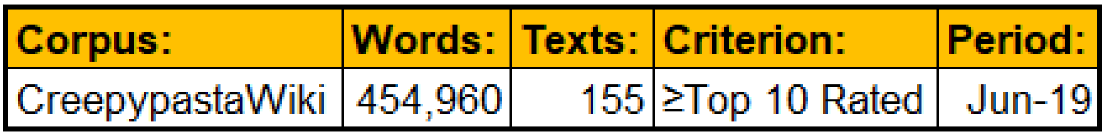
# Emerging horror

Exploring an archive of digitally borne horror narratives

*Abstract*

### Background

The objective was to explore if the regionality of folkloric narratives persists in digital material. This poster shows the early stages of scrutinizing a corpus of short stories appearing in the top-rated section of an online archive of user-submitted horror narratives, categorized with non-exclusive content tags by the users (Table 1).



| Corpus: | Words: | Texts: | Criterion: | Period: |
|---|---|---|---|---|
| CreepypastaWiki | 454,960 | 155 | ≥Top 10 Rated | Jun-19 |

Table 1: Corpus overview.

“Creepypasta” is the portmanteau of “Creepy” and “Copypasta”, referring to horror stories written to be repeatedly copy/pasted online (Blank & McNeill, 2018). Horror is influenced by the current fears and anxieties of both the author and the target audience; a feature that remains in the digital form (Bloom, 1998; Botting, 2008, 2013). The collaborative space allows for the spread and formulation of “Creepypasta” as a continuation of the urban legend (Victoria, 2018). Such stories are connected to the youth engaged in negotiating identity in different ways within those spaces (Scollon & Scollon, 2004).

Creepypasta themselves can be seen as the modern, digital folklore, suggested as “Netlore”, taking up the function of the earlier narrative traditions as interpretations of the surrounding world (Sanchez, 2019).The emergence of digital narratives includes the addition of new materials and features to netlore such as audiovisual components, a preference for first-person narration, faux-reality and the virtual anonymity of the author (Williams, 2015; Blank, 2007; Sanchez, 2019).

## Research Questions and Methods

Answering the question of regionally tied content in “Creepypasta” similarly to what is found in folklore and urban legends necessitates answering:

(1) Which places are frequently referred to in the story settings?

(2) What do the topics tell us about emergent digital horror in a region?

The NLTK © library in Python 3 © was used for querying the category tags and Recogito © was used for named-entity recognition and geographical mapping. The user-driven categorization and rating become central to this project, as the stories within the dataset have been selected due to their high rating within each of the 43 categories found in the archive. In order to provide an idea of frequent named environments within the stories, such as nations or cities, the Geonames coordinate dataset of populated areas was used for the mapping within Recogito ©.

## Geographical Locations

The Recogito © NER found 4,014 place names. While many of the occurrences are referring to larger geographical areas or nations or are drawn from adjectives the NER also found named cities (Figure 1).

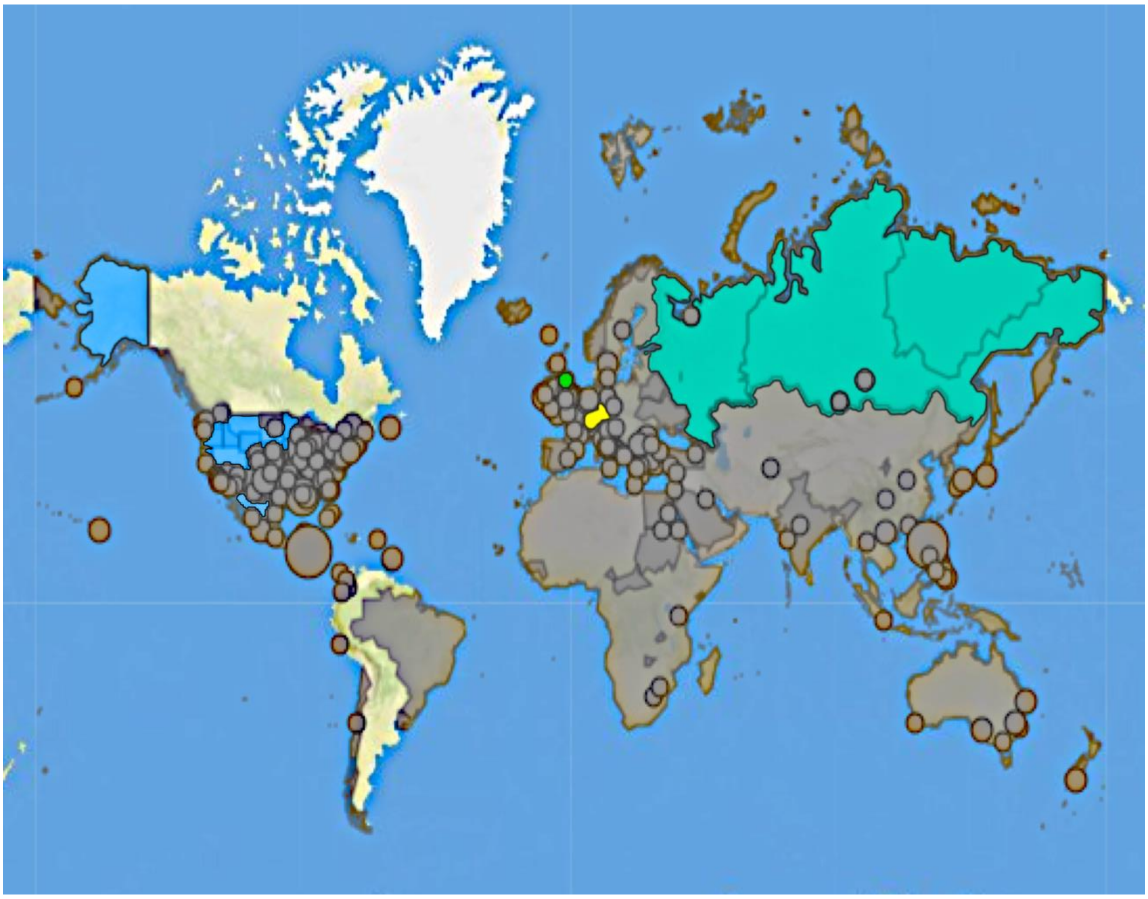


Figure 1: Named entities plotted by Recogito with the top 4 entities colored according to Figure 3.

The density of the coloration corresponds to the number of occurrences. The named regions are widespread, but the densest areas of named cities are found in the U.S.A. and the U.K., followed by Germany and France. Paris is the only city found in the top frequencies, followed by the U.S.A.

### Category Tags

Figure 2 shows the top-ranking of categories. While mental illnesses and historical settings are classical parts of the horror genre, the popularity of computers, the internet and audiovisual media is a new addition to the concept, although popular in contemporary horror films. This supports the idea of a Netlore emerging as a continuation of the folklore and urban legend concepts in horror; being linked to the scientific, cultural and speculative conventions of their time (Powell, 2018).

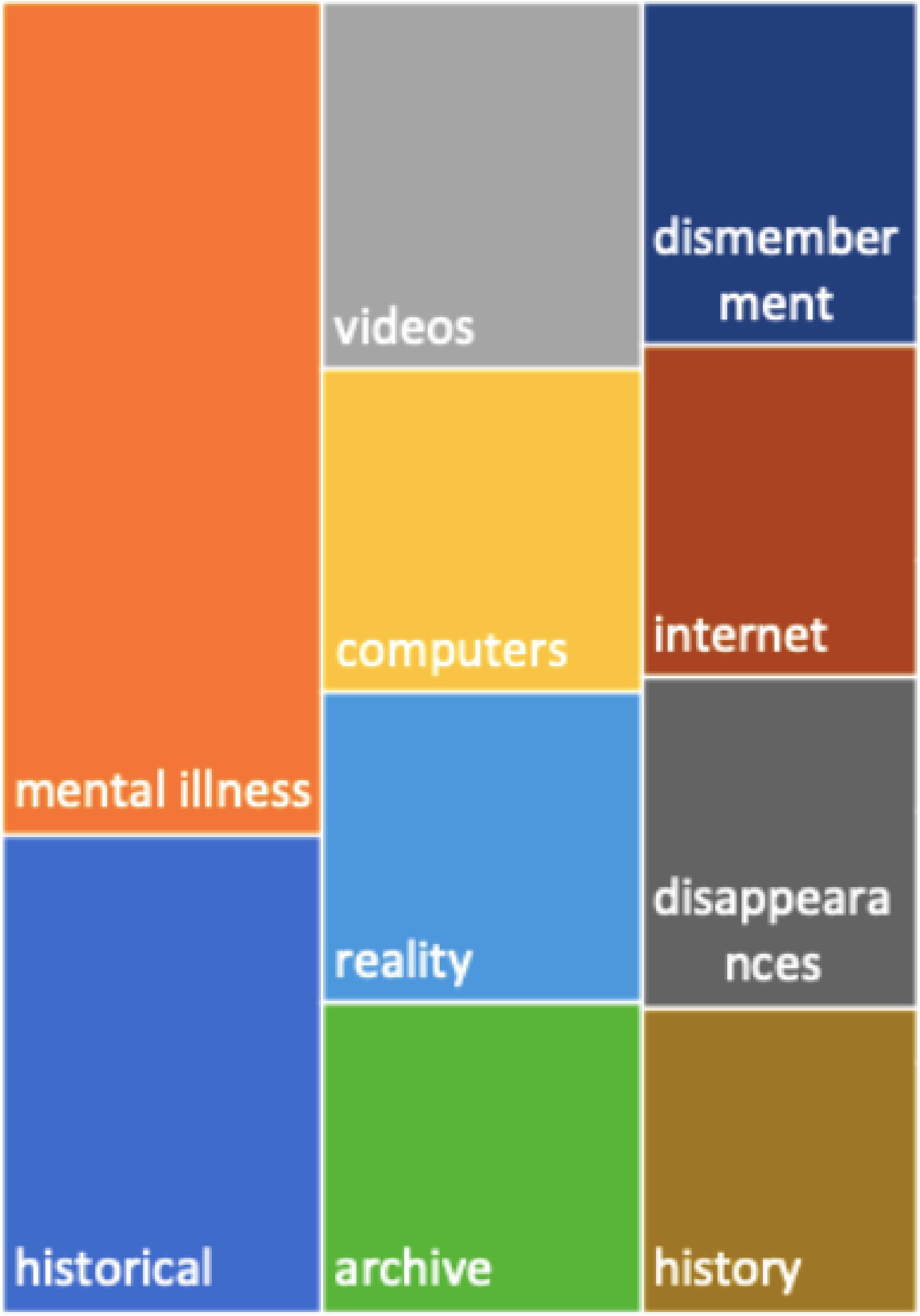


Figure 2: Top category tags in the dataset.

## Results

The most frequent locations found by the NER were the U.S.A., Russia, Paris and Germany. When splitting the dataset by location it the categories seem linked with the setting to some extent (Figure 3). The top topics (Figure 2) were found in most of the subsets of the data, but the popularity differed. Both the U.S.A. and Russia connotate “History”, while “Lovecraftian” is only found in the U.S.A. The popularity of history and historical events supports the idea of Netlore as a way for the digital generation explore both their own and the world’s history through legends, myths and oral tradition facilitated by new storyscapes (DeVos, 2012).

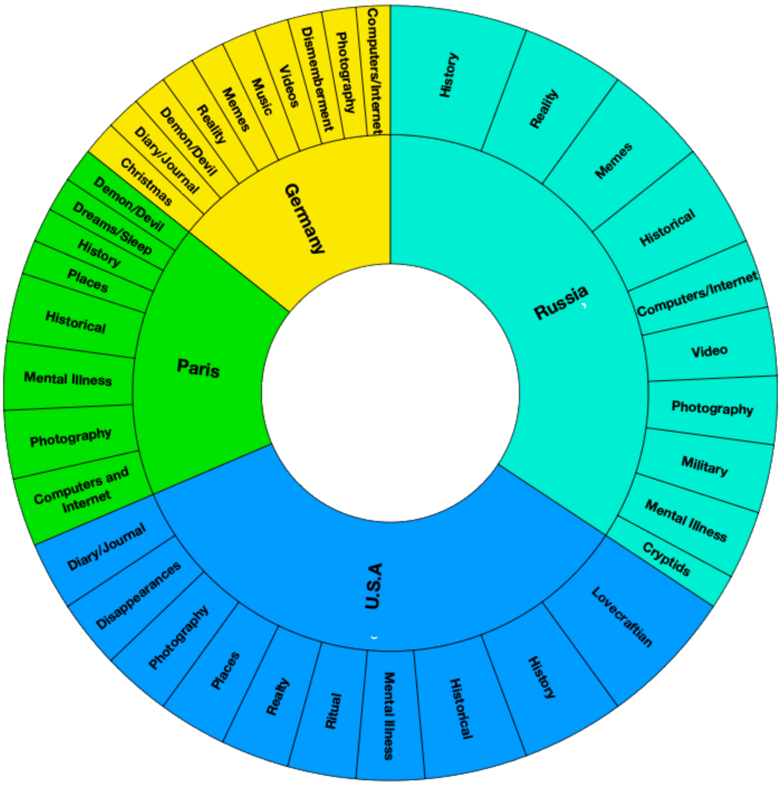


Figure 3: Category distribution by top 4 frequency named entities.

It seems that these narratives are the projection of urban legends in a new medium, reclaiming folklore in a digitized context. In that context, the places add value not to the topics in the stories but to their accounts (Frank, 2011). Combined with the faux-reality and first-person perspective suggested by previous material on the topic of netlore, the popularity of “Mental Illness” across the full dataset could be indicating mental welfare and well-being as existing fears amongst the digital generation.

“Photography” appears in all 4 NER categories, indicating the emergence of visual storytelling in the medium as narratives categorized with the tag by the community commonly contains pictures. The popularity of category tags like “Memes”, “Videos”, “Photography” and the previously mentioned “Computers and Internet” categories all support William’s (2015) description of Netlore as incorporating new features specific to the digital medium, while “Reality” would correspond to the genre as described by Sanchez (2019).

Future studies could explore these new features in relation to offline collections and in region-specific sub-corpora. Results could also be combined with the user ranking to determine patterns of preference based on location.

*References*

Blank, T. J. (2007). Examining the transmission of urban legends: Making the case for folklore fieldwork on the internet. *Folklore Forum, 3*(1), 15-26.

Blank, T. J., & McNeil, L. S. (2018). Slender man is coming: Creepypasta and contemporary legends on the internet. Colorado: Utah State University Press.

Bloom, C. (1998). *Gothic horror*. London: MacMillan Press.

Botting, F. (2008). *Gothic romanced*. Hoboken: Taylor and Francis.

Botting, F. (2013). *Limits of horror*. Manchester: Manchester University Press.

DeVos, G. (2012). *What happens next? Contemporary urban legends and popular culture*. Santa Barbara: Libraries Unlimited.

Frank, R. (2011). *Newslore: Contemporary folklore on the internet*. Mississippi: University Press of Mississippi.

Powell, D. W. (2018). *Horror culture in the new millennium: Digital dissonance and technohorror*. New York: Lexington Books.

Sánchez, S. (2019). Netlore: Urban legends and creepypastas. *DeSignis, 30*, 133-144.

Scollon, R., & Scollon, S. W. (2004). *Nexus analysis: Discourse and the emerging internet*. London: Routledge.

Victoria, E. A. (2018). Lovecraftian creepypastas: Digital cosmic horror. *Heresy and Beauty: Magazine of Cultural Studies on the Gothic Movement, 6*, 117-127.

Williams, C. (2015). How has creepypasta transformed folklore? *Digital Humanities Symposium*.

**Trapateau, Nicolas**

***Université de Nice, CNRS, BCL***

**nicolas.trapateau@univ-cotedazur.fr**

**Zumstein, Franck**

***Université de Paris, CLILLAC-ARP EA3967***

**franck.zumstein@u-paris.fr**

**Duchet, Jean-Louis**

***Université de Poitiers, FoReLLIS EA3816***

**jean-louis.duchet@univ-poitiers.fr**

Type of Contribution: Work-in-Progress Report

## Walker (1791) on the web

When hypertext serves diachronic phonology

*Abstract*

Inspired by pioneer works such as the *ECEP database*, our project aims at bringing to the fore the latest findings about English pronunciation in the still under-researched Late Modern English period (Yáñez-Bouza, 2020). Among the orthoepic dictionaries recording pronunciation at that time, Walker's *Critical Pronouncing Dictionary* (1791) provides one of the most elaborate pre-phonetic transcription systems and a verbose front matter containing 560 pronunciation principles. Out of its 39,000 dictionary entries, more than a thousand contain critical notes revealing phonetic details, variants and the sociolinguistic pressures of the time. The recent completion of a critical TEI edition of the dictionary facilitates querying for specific lexical sets or stress patterns as in any lexico-phonetic corpus. However, with the progressive enrichment of the xml file with annotations and the encoding of the original internal cross-references, there is a greater need for an enhanced consultation tool to better

visualize the data. Our work in progress is therefore to produce a web version of Walker's dictionary in order to get a global vision of its wealth of scattered but highly interconnected information.

Phonological data in Walker's dictionary can be found in three different locations: the pre-phonetic transcription of each entry word, the author's critical notes included in some of the entries, and the principles of pronunciation that appear in the front matter and to which some entries may refer. Our website takes advantage of that built-in hypertextual structure connecting orthoepic principles with the entries and vice-versa. In each entry, the reference to another entry is linked to that entry. Each principle referred to in an entry is accessible through a hyperlink, as is each reference to a principle in the text of another principle. An important added value of the web version is to provide, at the end of each principle, an additional symbol giving access to a list of the entries quoting it. Each of those list items is a link giving access to the full entry, which makes the cross-references totally bi-directional. Additional annotation is provided by a representation of the stress pattern and an equivalent IPA transcription of each entry to make the author's original discourse more readily interpretable and to allow further data mining beyond that discourse.

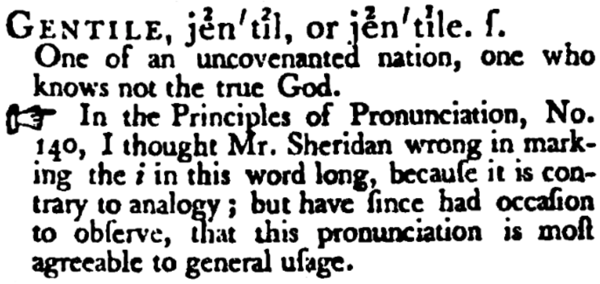
A list of the words with a variant spelling not easily retrievable in the alphabetical order of entries (e.g. *inforce*, *intire*, instead of *enforce*, *entire*) will provide the appropriate link to the entry in its original spelling and alphabetical order. "Walker on the Web" is a web site in which most pages are related to one another within a complex structure. To meet our needs, we have used *eXeLearning,* a free authoring tool, to produce the website. Though it is primarily intended for pedagogical resources, and particularly interactive content, it helps create XHTML or HTML5 pages within a predefined structure and to finally export the created content into a single folder, our website personalized by customizable CSS stylesheets.

In the course of this computerizing task, we were able to evaluate how consistent the lexicographer is in his principles and his transcriptions, and how this project can help us detect the phonological changes under way at the time and the changes which were to occur in the 19th and early 20th centuries, e.g. the pronunciation of words ending in *-ile*.

In contemporary US usage, the unstressed adjectival ending *-ile* is reduced to a syllabic l [ᵊl], while a long i (the diphthong [aɪ]) has prevailed in GB usage. An extraction of all the entry words ending in *-ile* in Walker's dictionary reveals that 92% of them are pronounced with a short [ɪ] (noted i$^2$) in the 18th century, of which the US pronunciation seems to be the direct reflex (Trapateau, 2017). A number of the entries extracted contain a cross-reference to principles 140 in the front matter, which formulates a pronunciation rule for the <i> based on rhythmic stress.

> 140. There is one rule of very great extent, in words of this termination, which have the accent on the penultimate syllable, and that is, that the *i* in the final syllable is short; thus *servile*, *hostile*, […] are pronounced as if written *sevil*, *hostil*, […]. *Myrrhine*, *vulpine*, and *gentile*, though marked with the *i* long by Mr. Sheridan, ought, in my opinion, to conform to the general rule, and be pronounced with the *i* short. (Walker, 1791)

The principle gives access both to the author's linguistic reasoning but also to variation recorded in the work of his contemporaries like Sheridan (1780). In fact, the short i is present in only 70% of Sheridan's *-ile* words, and as little as 38% in other authors like Buchanan (1766) (cf. Trapateau, 2017). Further information may be found in the critical notes of some word entries that may challenge the rules set in the principles, as in the entry for the adjective *gentile*, where Walker changes his mind.



GENTILE, j$\overset{2}{\text{e}}$n′t$\overset{2}{\text{i}}$l, or j$\overset{2}{\text{e}}$n′t$\overset{1}{\text{i}}$le. ſ.
One of an uncovenanted nation, one who knows not the true God.
☞ In the Principles of Pronunciation, No. 140, I thought Mr. Sheridan wrong in marking the *i* in this word long, becauſe it is contrary to analogy; but have ſince had occaſion to obſerve, that this pronunciation is moſt agreeable to general uſage.

The fact that the pronunciation with short i is placed first in the entry needs therefore to be revaluated against the discourse contained in both the critical note and the principle. That discourse is what makes Walker's dictionary stand out from many other eighteenth-century works that only provide silent lists of transcriptions. In many instances, where usage is unstable, it explicitly favours the pronunciations that follow analogy rather than exceptional realizations. This tendency needs to be taken into account to check for potential bias in the data, but it rarely leads to artificial pronunciations since variation is at least mentioned and widespread usage acknowledged (cf. *gentile*).

The principles also reveal implicitly that Walker's approach is often led by morphology. When comparing for instance the quantity of the i in all the Latin adjectives in *-ilis* and their English equivalent in *-ile,* Walker shows that the quantity of the penultimate i varies in Latin, whereas the pressure of analogy in English tends to choose only one pronunciation for an entire morphological class.

The web edition of Walker's dictionary not only improves the readability of the original material, documenting the author's improvements in all the editions he supervized, but also reconnects the pre-phonetic data with the discourse on that data. Such recontextualization is necessary to interpret accurately this important testimony for diachronic phonology and its role in the evolution of today's varieties of English.

## *References*

Buchanan, J. (1766)*. An essay towards establishing a standard for an elegant and uniform pronunciation of the English language*. London: Edward & Charles Dilly.

*Eighteenth-Century English Phonology Database* (ECEP). (2015). Compiled by J. C. Beal, N. Yáñez-Bouza, R. Sen, & C. Wallis. The University of Sheffield and Universidade de Vigo. Retrieved from: http://www.hrionline.ac.uk/ecep.

*eXeLearning*. (2019). Instituto de Tecnologías Educativas del Ministerio de Educación del Gobierno de España, Madrid, Spain. Retrieved from: https://exelearning.net/en/.

Sheridan, T. (1780). *A general dictionary of the English language*. London: J. Dodsley, C. Dilly & J. Wilkie.

Trapateau, N. (2017). Dating phonological change on the basis of eighteenth-century British English dictionaries and orthoepic treatises. *Dictionaries: Journal of the Dictionary Society of North America, 38*(2), 1-29.

Walker, J. (1791). *A critical pronouncing dictionary and expositor of the English language*. London: G. G. J., J. Robinson & T. Cadell.

Yáñez-Bouza, N. (2020). ECEP: Historical corpora, historical phonology and historical pronouncing dictionaries. *English Language and Linguistics*, 1-18.

**Valvason, Elena**

***University of Pavia & University of Bergamo***

**elena.valvason01@universitadipavia.it**

Type of Contribution: Full Paper

# There is urgency to achieve the sustainable development goals

## A corpus-assisted study of the discursive construction of sustainable development in the British press

*Abstract*

### Introduction

On 25th September 2015, the United Nations issued *The 2030 Agenda for Sustainable Development* (United Nations, 2015). The 2030 Agenda is a resolution intended to trace the path that the world's nations should follow to advance towards sustainable development. According to the UN, this path should consider the pivotal role of "people", "planet", "prosperity", "peace", and "partnership" and it should tackle seventeen Sustainable Development Goals (or SDGs) in order to be fulfilled (United Nations, 2015, p. 2).

Since the release of the Agenda, the commitment to the Sustainable Development Goals has involved governments, organizations, and citizens (Fox and Stoett, 2016). In addition, it has also stimulated much discussion in the media. Before the publication of the document, however, linguistic research had noticed that sometimes the multiword expression *sustainable development* was used vaguely in news discourse (Alexander, 2002). Furthermore, it showed that this multiword unit was discursively constructed mainly as a goal that governments and organizations wished to reach (Mahlberg, 2007).

Elaborating on this point, the present research aims at sketching the discursive construction of sustainable development in a sample of British news discourse appeared after the publication of the 2030 Agenda.

### Research questions

The goal of the paper is two-fold. On the one hand, the analysis aims to track the lexico-grammatical behaviour of *sustainable development* and *sustainable* in British news discourse through their collocational patterns. On the other, it seeks to fill the linguistic information gathered from the lexico-grammatical behaviour of the two lexical items with a glance at the lexical preferences of the selected newspaper articles in terms of keywords.

**Approach**

The study is carried out within the realm of the ecological analysis of discourse (cf. Alexander & Stibbe, 2014; Stibbe, 2015) and with a corpus-assisted approach (cf. Partington et al., 2013; Marchi & Taylor, 2018). In other terms, a corpus-aided methodology enables to quantitatively spot and strengthen linguistic observations that expose discourse that might have a beneficial or detrimental impact on the relationship between the entities residing in nature (Alexander, 2018; Poole, 2019).

**Data**

The study is conducted on a corpus of 500 newspaper articles published between 2016 and 2018 in the most read British quality papers (i.e. *Financial Times*, *The Daily Telegraph*, *The Guardian*, *The Times*). The articles were collected on the LexisNexis repository with the query expression *sustainable development* and then sampled to reach the overall number of 500. The corpus counts 432,736 running tokens and it is divided into four subcorpora. The *Financial Times* subcorpus includes 79 articles and 62,580 tokens; *The Daily Telegraph* subcorpus is made of 46 articles and 33,931 words; *The Guardian* subcorpus consists of 301 articles and 281,567 tokens; *The Times* subcorpus is made of 74 articles and 54,658 tokens.

**Method**

First, the corpus is explored with the #LancsBox software (Brezina et al., 2018) in search for the collocational behaviour of *sustainable development* and *sustainable*. Collocational patterns are extracted first for the whole corpus and then for the four subcorpora separately. The collocates are searched for within a span of 5 words to the left and 5 words to the right of the node using the Z statistical measure with a statistical threshold of 10.0 and a frequency threshold of 5. The collocational networks of the lexical items are analyzed in terms of their syntactic and semantic features. They are grouped according to semantic areas and connotation to highlight their semantic preference (Sinclair, 1991) and their semantic prosody (Louw, 1993). In addition, their concordance lines are studied in depth thanks to the AntConc software (Anthony, 2019).

Second, the corpus is analyzed through the SketchEngine platform (Kilgarriff et al., 2014) to extract its keywords. The keywords are calculated by comparing the collection of newspaper articles on sustainable development to the English component of the *Timestamped JSI webcorpus 2014-2019*, namely a continuously updated collection of newsfeed (Bušta et al., 2017). The keywords are computed with the simple maths score set at 100 for rarity and at 1 for minimum frequency. The highest-ranking keywords are then grouped according to their matching one or more of the seventeen SDGs.

## Results

The analysis of the collocational networks shows that in this corpus *sustainable development* is generally constructed in relation to the 2030 Agenda and to its goals and targets (thanks to collocates like *agenda*, *goals*, *2030*, *17*, etc.). While this is common to all newspapers, *The Guardian* proves richer in collocates. Its collocational network adds a set of material processes that depict *sustainable development* as a positive goal to be reached (e.g. *achieve*, *achieving*, *implementation*, etc.). The only collocates hinting at the fields in which sustainability should be aimed for are *business* and *climate*.

The former, in particular, introduces one of the semantic areas that the immediate collocates of the adjective *sustainable* can be grouped into. In most subcorpora, in fact, *sustainable* modifies lexemes like *growth*, *investing* and *investment*. These lexical items belong to the semantic field of economy and they are the second most significant co-occurring lexemes for of the adjective. The strongest collocate of *sustainable*, in fact, is *development*, due to the nature of the corpus. After the noun *development* and the lexemes hinting at economic matters, the lists of collocates feature lexical items referring to the environment (e.g. *environment*, *palm oil*, and *water*). However, this semantic field is represented only in *The Guardian* subcorpus and it does not seem to raise a consistent and robust engagement with environmental problems.

The study of the keywords of the whole corpus in comparison with the *Timestamped JSI webcorpus 2014-2019* highlighted lexemes referring to sustainable development in general (e.g. *development, sustainable, sustainability*) and to the 2030 Agenda (e.g. *goals, SGDs, UN*). Then it stressed the engagement of the newspaper articles for the Sustainable Development Goals explicitly dealing with poverty (*poverty*), health (*health* and *TB*), education (*education*), water and sanitation (*water*), energy (*energy* and *green*), and climate change (*carbon, change, climate, emissions, Paris*, and *pollution*). Although the range of semantic areas covered by the lexemes in the keyword list is wider compared to the one of the collocational networks and the environmental issues are represented better, it seems that the well-being of nature is still partially neglected by the newspaper articles.

Overall, this analysis of British news discourse on sustainable development shows that, also after the release of *The 2030 Agenda for Sustainable Development*, sustainable development is discursively constructed as a goal to aimed at by tackling economic and environmental issues and by dealing with various aspects of human well-being.

## *References*

Alexander, R. (2002). Everyone is talking about ‘sustainable development’. Can they all mean the same thing? Computer discourse analysis of ecological texts. In F. Alwin, H. Penz & W. Trampe (Eds.), *Colourful green ideas: Papers from the conference 30 Years of Language and Ecology* (pp. 239-254). Bern: Peter Lang.

Alexander, R. (2018). Investigating texts about environmental degradation using critical discourse analysis and corpus linguistics techniques. In F. Alwin & H. Penz (Eds.), *The Routledge handbook of ecolinguistics* (pp. 196-210). New York/London: Routledge.

Alexander, R., & Stibbe, A. (2014). From the analysis of ecological discourse to the ecological analysis of discourse. *Language Sciences, 41*, 104-110.

Anthony, L. (2019). *AntConc (Version 3.5.8)* [Computer Software]. Retrieved from http://www.laurenceanthony.net/software.

Brezina, V., Timperely, M., & McEnery, T. (2018). #LancsBox v. 4.x [Computer Software]. Retrieved from http://corpora.lancs.ac.uk/lancsbox.

Bušta, J., Herman, O., Jakubíček, M., Krek, S., & Novak, B. (2017). JSI Newsfeed Corpus. Paper presented at *9th International Corpus Linguistics Conference*. Birmingham, 25-28 July 2017.

Fox, O., & Stoett, P. (2016). Citizen participation in the UN sustainable development goals consultation process: Toward global democratic governance? *Global Governance, 22*(4), 555-573.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: 10 years on. *Lexicography, 1*, 7-36.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157-176). Philadephia/Amsterdam: John Benjamins.

Mahlberg, M. (2007). Lexical items in discourse: identifying local textual functions of *sustainable development*. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Eds.), *Text, discourse and corpora: Theory and analysis* (pp. 191-218). London/New York: Continuum.

Marchi, A., & Taylor, C. (2018). Introduction: Partiality and reflexivity. In T. Charlotte & A. Marchi (Eds.), *Corpus approaches to discourse* (pp. 1-15). London/New York: Routledge.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam/Philadelphia: John Benjamins.

Poole, R. (2019). Expanding the scope of corpus-aided ecolinguistics. Paper presented at *4th International Conference on Ecolinguistics* (University of Southern Denmark, Denmark), 12-15 August 2019.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stibbe, A. (2015). *Ecolinguistics*. London/New York: Routledge.

United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development*. Retrieved from: https://sustainabledevelopment.un.org/post2015/transformingourworld.

**Van Vuuren, Sanne**

***Radboud University***

**s.v.vuuren@let.ru.nl**

**Berns, Janine**

***Radboud University***

**j.berns@let.ru.nl**

**Bank, Marketa**

***Radboud University***

Type of Contribution: Full Paper

# Syntactic complexity and L1 influence in nominal postmodification

*Abstract*

**Background**

Syntactic and pragmatic cross-linguistic differences make (pro)nominal postmodification patterns an interesting object of study, not only for typologists, but also for language acquisition researchers. English has two devices for clausal postmodification: relative and participle clauses, which both allow the speaker to further specify the noun without having to start a new sentence, cf. (1). In English, present and past participle clauses can be used as alternatives for relative clauses (Huddleston et al., 2002: Sleeman, 2017), serving to condense the message without affecting its clarity:

(1) A bike was stolen from his shed. The bike/It was spotted outside a chip shop.

(2) a. The bike that was stolen from his shed was spotted outside a chip shop. [relative clause]

b. The bike stolen from his shed was spotted outside a chip shop.
[past participle clause]
c. The man walking to work is my boss.
[present participle clause]

Existing research has reported that EFL learners from different L1 backgrounds use fewer non-finite clauses than native speakers (NSs) (e.g. Granger, 1997; Parrot, 2000; Hinkel, 2002), and learners thus seem less inclined to produce a sentence such as (2c). In this study, we further explore the roles of transfer and linguistic maturity (e.g. Yang, 2013; Parkinson & Musgrave, 2014) in the postmodification patterns produced by EFL learners by comparing argumentative texts written by learners from three typologically distinct L1 backgrounds, i.e. Czech, Dutch and French.

Czech, French and Dutch each possess structures that can be considered equivalent to English relative and participle clauses, although they don't have bare relatives. Other differences reside in the frequency of use, stylistic and pragmatic restrictions on the use of participle clauses and preferences for pre- or postmodification:

| | |
|---|---|
| **Dutch** | ✗ Participles (especially present participles) mostly restricted to formal language, fixed expressions and narrative language (e.g. Aarts & Wekker, 1987: 146-147; De Moor & Copriau, 1998: 309). Often used prenominally. |
| **French** | ✗ Present participles more common in formal language (e.g. Escoubas-Benveniste et al., 2012). French adnominal participles generally postmodify the noun they relate to. |
| **Czech** | ✗ Czech deverbal adjectives function as equivalents of participle clauses and can be used for postmodification. Malá & Šaldová (2015) report 40% of English participle clauses are translated by deverbal adjectives in their translation corpus). |

## Method

Adnominal participle clauses and relative clauses were extracted from a subsection of the French, Dutch and Czech subcorpora of the *International*

*Corpus of Learner English* (ICLE, 2002)[1] and the *Louvain Corpus of Native English Essays* (LOCNESS, Granger, 1998). The corpora were parsed by means of the Stanford Parser (Klein & Manning, 2003), after which we identified adnominal participle and relative clauses using the *Corpus Editor for Syntactically Annotated Resources* (CESAR, Hoek & Komen, submitted). All output generated by the queries was manually checked.

## Results

Of the three learner groups, the Czech learners use the least postmodification and the Dutch learners the most. Both the NSs and the Dutch learners use significantly more clausal postmodification than either the French or the Czech learners.

All learner groups use significantly fewer participle clauses than the NSs. This includes both present and past participles. Although the French learners use slightly more present and past participles than the Dutch or the Czech learners (Figure 1), the differences between the learner groups for both types of participle clause are not significant.

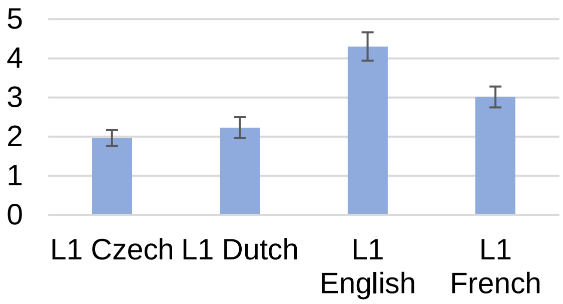


Figure 1: Participles per 1000 words.

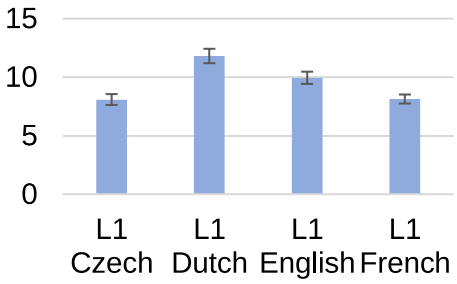


Figure 2: Relative clauses per 1000 words.

For the relative clauses (Figure 2), unlike the participle clauses, it is the French and the Czech learners that pattern together. Their use of bare relatives as well as relatives introduced with a relativizer is significantly lower than that of the

[1] The student contributors to the ICLE corpus have been classified as upper-intermediate to advanced on the basis of institutional status rather than an objective standardized proficiency test (cf. Granger et al., 2009: 12). As differing levels of proficiency might clearly affect our results, we used Lu’s syntactic complexity analyzer (cf. Lu & Ai, 2015) to calculate a number of interrelated global syntactic complexity measures that are considered to be reliable regardless of instructional setting (Ortega, 2003). These measures suggest that the Czech learners may be less advanced than the other two L1 groups.

NSs and Dutch EFL learners. The Dutch learners actually use significantly more relatives introduced by a relativizer than the NSs. Although bare relatives exist in none of the L1s represented in the learner corpus, they are only underrepresented in the writing of the French and Czech learners.

## Discussion and conclusion

All learners use significantly less non-finite clausal postmodification than the NSs. That does not automatically imply, however, that they have a stronger preference for finite, relative clauses compared with NSs. In fact, any type of clausal postmodification is underrepresented in the writing of the French and the Czech learners. The Dutch learners compensate their relatively lower use of participle clauses with a higher use of relative clauses, arriving at a similar rate of clausal postmodification as the NSs.

Our results so far suggest a combined effect of transfer and syntactic development. Dutch learners' preference for relative clauses - the least marked of both options for clausal postmodification in their L1 - seems likely to be transfer-related. This remains to be confirmed by future research examining comparable texts written by EFL learners in their L1. Compared with the native speakers, French learners use significantly fewer participle and relative clauses, but the relative contribution of each category to the total number of clausal postmodifiers is actually very similar to that of the NSs. The question is whether their rate of clausal postmodification would increase with increasing proficiency and whether their proportional use of participle and relative clauses would remain stable over time. The same goes for the Czech learners, whose syntactic complexity measures remain behind those of the other learners, suggesting that their lower use of clausal postmodification could be due to their relatively lower proficiency level at the time of data-collection. A longitudinal study tracing learners' use of relative and participle clauses over time would help to disentangle the relative contribution of L1 influence and syntactic maturity in clausal postmodification.

### *References*

Aarts, F. G. A. M., & Wekker, H. C. (1987). *A contrastive grammar of English and Dutch/Contrastieve grammatica Engels/Nederlands.* Dordrecht: Springer.

Cosme, C. (2008). Participle clauses in learner English: The role of transfer. In G. Gilquin, S. Papp & M. Belén Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 177-198). Amsterdam: Brill Rodopi.

De Moor, W., & Copriau, E. (1998). *A contrastive reference grammar*. Kapellen: Pelkmans.

Escoubas-Benveniste, M. P., Floquet, O., & Bolasco, S. (2012). Contribution empirique à l'étude du gérondif et du participe présent en français parlé et écrit. In A. Dister, D. Longrée & G. Purnelle (Eds.), *JADT 2012: 11èmes Journées internationales d'Analyse statistique des Données Textuelles* (pp. 473-485). Actes du colloque JADT 2012, Liège.

Granger, S. (1997). On identifying the syntactic and discourse features of participle clauses in academic English: Native and non-native writers compared. *Studies in English Language and Teaching*, 185-198.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). London/New York: Addison Wesley Longman.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *The International Corpus of Learner English.Handbook and CD-ROM* (Version 2).

Hinkel, E. (Ed.) (2002). *Second language writers' tekst*. London: Erlbaum.

Hoek, J., & Komen, E. (forthcoming). Quantitative corpus research for non-programmers.

Huddleston, R. D., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

*ICLE/International Corpus of Learner English*. (2002). Granger, S., Dagneaux, E., & Meunier, F. (Eds.). Louvain-la-Neuve: Presses Universitaires de Louvain.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics Volume 1* (pp. 423-430). Association for Computational Linguistics.

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16-27.

Malá, M., & Šaldová, P. (2015). English non-finite participial clauses as seen through their Czech counterparts. *Nordic Journal of English Studies, 14*(1), 232-257.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics*, *24*(4), 492-518.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, *14*, 48-59.

Parrot, M. (Ed.) (2000). *Grammar for English language teachers: With exercises and a key.* Cambridge: Cambridge University Press.

Sleeman, P. (2017). Participial relative clauses. In M. Aronoff (Ed.), *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press.

Yang, W. (2013). Response to Biber, Gray, and Poonpon (2011). *Tesol Quarterly*, *47*(1), 187-191.

**Wasserman, Ronel**

***North-West University***

**Ronel.Wasserman@nwu.ac.za**

Type of Contribution: Full Paper

## When *no* means *yes*

The grammaticalization of no in South African English

*Abstract*

This paper provides corpus evidence of the affirmative use of *no* in South African English (SAfE) and argues that longstanding contact with Afrikaans is the activating force of the grammaticalization of *no* in SAfE. The uniqueness of *no* in SAfE is often briefly mentioned and ascribed to Afrikaans influence in popular and academic literature (e.g. Bowerman, 2004; Branford & Venter, 2016;) but no corpus evidence has yet been reported. According to Branford and Venter (2016, p. 10) *no* is regularly used in SAfE to answer a question in the affirmative, for example if a person is asked whether they enjoyed a movie they might answer, ‘No, it was very good’ (cf. Silber 1991), or it can be used as an affirmative sentence initiator, e.g. if asked ‘How are you?’ the response ‘No, I’m fine, thanks’ is very common. Trudgill & Hannah (2008, p. 35) calls this the ‘non-negative’ use of *no* (a more tentative take on ‘affirmative’), especially as an introductory particle, where its function could be to “negate assumptions made in the preceding question or comment”.

In this study the use of *no* is analyzed in the spoken component of ICE-SA (Jeffery, 2003), which consists of 402,175 words. Concordance lists for the lemma *no* have been extracted via Wordsmith (version 7.0) and manually analyzed. No instances of affirmative *no* are attested in written corpora of SAfE, including the Historical Corpus of South African English (HICSAE) or the written component of ICE-SA. 51 instances of affirmative *no* were identified among the 2,123 tokens of *no* in the corpus, which means it accounts for 2.4% of uses of *no*. Among the semantic features of spoken SAfE, affirmative *no* is the third

most frequent, after emphasizer *now* and weak obligation *must* (cf. Jeffery & Van Rooy, 2004; Wasserman, 2014; 2019). This use is only detected in informal direct conversation, with the exception of one instance used in a classroom lesson.

From my analyses in spoken ICE-SA, the instances of *no* as a non-negative sentence initiator did however not adhere to Trudgill and Hannah's (2008) description of the function to negate previous assumptions. Instead, upon deeper analysis, *no* is firstly used either as a polite discourse marker, possibly even conveying mild surprise at a question, similar to the gambit *oh* - in a way assuring the interlocutor of something - as in (1), affirming the assumption that the boys are well, rather than dismissing or negating the assumption that they are not.

(1) M: … and now how're the boys?
S: **No** they're fine // Bruce has been writing exams so // now he's finished. (1990s; Direct conversation)

Logically the question 'How are you' or 'How are the boys', for example, does not contain a negative assumption, especially in South Africa, where it is most often asked as part of social ritual or politeness than out of real concern for a person's well-being, with some sort of negative assumption underpinning it (cf. Branford & Venter, 2016, p. 9).

In (2) *no* emphasizes that the speaker agrees with the preceding statement, with a layer of delight, which in this use would not be qualitatively different from either *oh yes* or *yes*, as Bowerman (2004, p. 479) claims.

(2) M: Show her // have you seen that thing?
S: Oh you've got that // oh yes // yes
N: Absolutely
S: Oh **no** they're very good // **no** they're marvellous (1990s; Direct conversation)

In the second kind of usage observed in the analysis, *no* is also not qualitatively different from *yes*, as seen in (3). The speaker in this example is affirming (and not negating) a previous negative statement. By contrast, in GenE "*yes* has been the ordinary affirmative response word in reply to any question positive or negative" since about 1600 (OED, 2018), making yes the unmarked option to affirm a negative statement, but SAfE chooses the Afrikaans-like alternative:

(3) A: Oh well nothing's easy in academics these days
X: (laugh) That's right
A: (laugh) Not a not a job for the faint-hearted
X: **No** that's right, uh, the the matter of the doctorate
A: uhm uhm (1990s; Direct conversation)

An additional layer of politeness is added in (4), reciprocating the hedging in negative *wouldn't*, but *no* is still clearly an answer in the affirmative, closely resembling the purely affirmative examples that follow.

(4) B: Wouldn't you like to go and have a rest man? // Go and lie down on the bed
C: **No** I will thanks (1990s; Direct conversation)

Thirdly, it does however also often happen that *no* in SAfE is plainly used to mean *yes*, either as a direct affirmative answer to a question, as in (5), or as a statement of agreement, as in (6). *No* in SAfE has therefore become quite bleached, losing much of its force in speech.

(5) A: Was it done on the same basis?
S: **No** I um it was done on the same basis but I can't remember how much tax came off. (1990s; Direct conversation).

(6) X: Ja that's what I'm saying
A: Ah great
X: Oh yes ja
A: Okay **no** that's good (1990s; Direct conversation)

These unique uses echo the semantic flexibility of Afrikaans cognate *nee* to venture into positive, affirmative territory, especially as used in informal, spoken contexts. The *Woordeboek van die Afrikaanse Taal* (WAT) (2018) and Van Jaarsveld (1991) list many examples where *nee* is used sentence-initially in Afrikaans, corresponding to the uses that were adopted in SAfE. Sentence-initial *nee* communicates three basic affirmative meanings. Firstly, it serves as a polite discourse marker very similar to the use in (1), secondly, as a positive emotive interjection, usually as a marker of surprise similar to (2), and lastly, *nee* is often used very clearly in the exact same sense as *ja* ('yes') to answer a question in the affirmative and to express agreement, which reflect the uses in (5) and (6) (Van Jaarsveld, 1991, pp. 80-81; WAT, 2018), making the case for *nee*'s influence very strong.

The conclusion is firstly drawn that contact-induced grammaticalization has occurred based on structural resemblance and without any language-internal propensity for *no* to extend (and even reanalyze) its semantic domain from negative to affirmative (cf. Heine & Kuteva, 2005). Secondly, in this case Afrikaans is not only the source language for a calque on *nee*, but acts as catalyst in activating the grammaticalization process in the semantic extension of *no*, an existing word (construction) in English. Lastly, affirmative *no* is found to be a purely spoken feature of SAfE in interactive contexts.

### *References*

Bowerman, S. (2004). White South African English: Morphology and syntax. In B. Kortmann, E.W. Schneider, K. Burrage, R. Mesthrie & C. Upton (Eds.), *A handbook of varieties of English: Morphology and syntax* (pp. 472-478). Berlin/ New York: Mouton de Gruyter.

Branford, J., & Venter, M. (2016). *Say again? The other side of South African English*. Cape Town: Pharos.

Heine, B., & Kuteva, T. (2005). *Language contact and grammatical change*. Cambridge: Cambridge University Press.

Jeffery, C. (2003). On compiling a corpus of South African English. *Southern African Linguistics and Applied Language Studies, 21*(4), 341-344.

Jeffery, C., & van Rooy, B. (2004). Emphasiser now in colloquial South African English. *World Englishes, 23*(2), 269-280.

Oxford English Dictionary (OED). (2018). No. In *Oxford English Dictionary (Online)*. Retrieved from http://www.oed.com.nwulib.nwu.ac.za.

Silber, G. (1991). *It takes two to toyi-toyi: A survival guide to the New South Africa.* Johannesburg/ New York: Penguin.

Trudgill, P., & Hannah, J. (2008). *International English: A guide to the varieties of Standard English.* 5th edition. New York: Routledge.

Van Jaarsveld, G. J. (1991). Oor ja, nee en ja-nee in Afrikaans. *Acta Academica, 23*(1), 68-84.

Wasserman, Ronel. (2014). *Modality on trek: Diachronic changes in written South African English across text and context*. [Doctoral thesis, North-West University]. NWU Repository.

Wasserman, R. (2019). Historical development of South African English: semantic features. In R. Hickey (Ed.), *English in multilingual South Africa: The linguistics of contact and change* (pp. 52-73). Cambridge: Cambridge University Press.

Woordeboek van die Afrikaanse Taal (WAT). (2018). Nee. In *Woordeboek van die Afrikaanse Taal (Elektroniese weergawe)*. Laaste gebruik 30 April 2018. Retrieved from: http://www.woordeboek.co.za.nwulib.nwu.ac.za.

**Weilinghoff, Andreas**

***TU Dortmund University***

**andreas.weilinghoff@udo.edu**

Type of Contribution: Full Paper

# Aitken's law revised

## Vowel length in 21st century Scotland

*Abstract*

One of the most notable features of Scottish English is the Scottish Vowel Length Rule (SVLR) / Aitken's Law. It states that vowels are lengthened in stressed open syllables, when followed by voiced fricatives, /r/ and morpheme boundaries. Vowel durations are short in all other contexts which partially contradicts the Voicing Effect (VE), a lengthening pattern found in many varieties of English. Whereas the SVLR once operated amongst almost all vowels and diphthongs in Scottish English, newer studies have shown that its effects can be detected in three vowels (/i:/, /u:/, /ai/) only. Yet, no research project has investigated the vowel length patterns in connected Scottish Standard English (SSE) speech and many findings of previous studies remain contradictory. The unresolved questions of the geographical scope of the SVLR as well as the influence of age and gender-related variation have not yet been satisfactorily answered.

The present research paper addresses these questions. It incorporates a large-scale investigation into the SVLR analyzing the long high vowels /i:/ and /u:/ in stressed position of both monosyllabic and polysyllabic words. The study incorporates an up-to-date dataset with a large number of informants (N=102) from different age groups, genders and four dialect regions of Scotland (Northern, Highland, West Central and East Central). Unlike most previous investigations, which elicited speech in controlled laboratory settings, the present study analyzes the effects of Aitken's Law in naturally occurring language (SSE) and takes prosodic factors into account. The major source is

the yet unpublished Scottish component of the ICE corpora (Schützler, Gut & Fuchs, 2017), which includes time-aligned and manually corrected phonemic transcriptions of the spoken data.

The vowel durations as well as information about the surrounding environments were extracted with the help of a Praat script. In the next step, based on the list provided by Roach (2009), all function words were excluded from the dataset as they are frequently reduced in connected speech. Furthermore, personal names, place names, acronyms as well as all hits with the target vowel in unstressed position were excluded from the dataset as well. The final token number totals N=5562, with 3684 tokens for /i:/ and 1978 tokens for /u:/. Finally, further sociolinguistic and prosodic background information was manually added to each token. The categorization and labelling was based on the approaches by Chevalier (2019) and Rathcke and Stuart-Smith (2016). Statistical testing was carried out by means of linear mixed-effects models in R (version 3.6.1) with the *lme4* and *lmerTest* packages. The dependent variable was vowel duration in milliseconds (*durms*), fixed factors include *vowel*, *SVLR environment*, *gender*, *age group*, *dialect area* as well as *text type*, *speech form* and *phrasal position*. The variables *speaker* and *word* were treated as random factors. Non-significant variables were excluded from the model after multiple runs and post hoc testing was conducted by means of Tukey's HSD test.

The results show that the SVLR operates in the high vowels of 21st century SSE. Yet, also a small influence of the VE could be detected in the dataset. In general, prosodic factors, such as phrasal position, have a strong influence on vowel lengthening patterns in Scottish English. Therefore, the study provides further evidence for the claim that context dependent vowel length differences are smaller in spontaneous connected speech than in utterances elicited in a laboratory setting (Tanner et al., 2019). Age-related variation could only be detected for the *age group old*. That is to say that SVLR lengthening patterns apply more consistently in the speech of older speakers. The influence of the dialect regions as well as gender-related variation, however, turned out to be insignificant.

## *References*

Chevalier, F. (2019). On sound change and gender: The case of vowel length variation in Scottish English. *Anglophonia* 27.

Rathcke, T. V., & Stuart-Smith, J. H. (2016). On the tail of the Scottish vowel length rule in Glasgow. *Language and Speech, 59*(3), 404-430.

Roach, P. (2009). English phonetics and phonology: A practical course. Cambridge: Cambridge University Press.

Schützler, O., Gut, U., & Fuchs, R. (2017). New perspectives on Scottish Standard English: Introducing the Scottish component of the International Corpus of English. In S. Hancil & J. C. Beal (Eds.), *Perspectives on Northern Englishes* (pp. 273-302). Berlin/Boston: Mouton de Gruyter.

Tanner, J., Sonderegger, M., Stuart-Smith, J., & The SPADE Consortium. (2019). Vowel duration and the voicing effect across dialects of English. *Toronto Working Papers in Linguistics, 41*, 1-13.

**Yanagi, Tomohiro**

***Chubu University***

**yanagi@isc.chubu.ac.jp**

Type of Contribution: Poster Presentation

## Negative indefinites as negative polarity items in negative sentences

*Abstract*

This presentation argues that Negative Indefinites (NIs) such as *nænig* 'not-any' could function as Negative Polarity Items (NPIs) such as *ænig* 'any' in Old English (OE). Both types of lexical items were used in negative sentences introduced by the negative particle *ne* 'not' or *ne*-contracted forms like *næfde* 'not-had' (Ingham, 2006; Mitchell, 1985). Examples of NIs and NPIs are given In (1) and (2), respectively. In these examples, the negative particle and *ne*-contracted form are italicized and the objects with the NPI and NI are in boldface.

(1) þæt he næfre **nænige godcunde englas** *næfde*
'that he never had any god-like angels'
(LS 32 181.186)

(2) Forðon *ne* meaht ðu me nu ofer ðisne dæg **ænige helpe ne geoce** gefremman
'for after this day you cannot give me any help or aid'
(Bede 14.438.16)

On the basis of data retrieved from the York-Toronto-Helsinki parsed corpus of Old English prose (Taylor et al., 2003; YCOE), I discuss syntactic similarities and differences of NPIs and NIs in OE.

Semantically, although (1) contains the three negative elements (*næfre* 'never', *nænige* 'not-any' and *næfde* 'not-had'), the sentence is interpreted as negative. This phenomenon is called Negative Concord (Hogg, 2002; Mitchell, 1985; Wallage, 2017). By contrast, only the single negative element *ne* is used

in the negative sentence containing the NPI *ænig* ‘any’ in (2) (*ne* in boldface corresponds to the conjunction *nor* in Present-day English).

It is generally said that in OE, negative objects tend to precede lexical verbs (Pintzuk & Taylor, 2006; Ringe & Taylor, 2014, among others). Pintzuk & Taylor (2006) show that negative objects precede verbs (83.3%) more frequently than quantified objects, i.e. objects with quantifiers such *eall* ‘all’, precede verbs (59.9%), over the whole period of OE. If the OE period is divided into two subperiods, early OE (before 950) and late OE (after 950), the frequency of each object type will be slightly higher in early OE: the frequency of the ‘negative-object-verb’ order is 91.8% and that of the ‘quantified-object-verb’ order is 63.5%. In late OE, on the other hand, the frequency of the negative-object-verb order is 78.3% and that of the quantified-object-verb order is 56,4%. The four types of word order patterns are given in (3) and (4). The ‘object-verb’ and ‘verb-object’ order of each object type are illustrated in the (a) and (b) examples, respectively. (3) is of negative objects, and (4) quantified objects. In these examples, the lexical verbs are italicized and the objects are in boldface.

(3) a. þæt he ne mæge **nan god** *don*
‘that he can do no good’
(ÆLS (Memory of Saints) 295/Pintzuk & Taylor, 2006, 258)
b. ðæt he nolde *habban* **nane gemodsumnesse** wið ða yfelan
‘that he would have no agreement with the wicked’
(CP 46.353.2/ibid.)

(4) a. hu heo ana mihte **ealle þa gewytan** *awægan* mid aðe
‘how she alone could deceive all the sages with an oath’
(ÆLS (Eugenia 223/ibid.)
b. þe hæfde *geinnod* **ealle þas halgan**
‘who had lodged all the saints’
(ÆLS (Sebastian) 382/ibid.)

Through a corpus search of the YCOE, I consider the distribution of objects with *nænig* ‘not-any’ and their governing verbs. The results show that NI objects are more likely to precede their governing verbs in negative sentences, as in (1). Although NI objects contain the negative element, their distribution is close to that of quantified objects rather than that of negative objects. The frequency of

NI objects preceding lexical verbs is 65.0%. It is similar to that of quantified object (63.5%), but not that of negative objects (91.8%). Given that the NI *nænig* 'not-any' was mainly used in early OE, it can be concluded that the distribution of NI objects and quantified objects are quite similar, with respect to the object-verb word order patterns.

Moreover, objects with NPIs precede their governing verbs in negative sentences, as in (2), with almost the same frequency of NIs, i.e. 62.5%. From these facts of NIs and NPIs, it can be concluded that NIs functioned as NPIs in negative sentences of OE.

I also argue that in OE, NIs could occur in the subject position, as in (5), more frequently than NPIs and that if an NPI occurred in the subject position, there was a tendency that the NPI followed the negative particle *ne*, as in (6).

(5) **nænig** Sceotta cyninga *ne* dorste [. . .] cuman oð ðysne andweardan dæg
'no king of the Scots dare come up to this present day'
(Bede 1 18.92.24)

(6) *Ne* **ænig** man oðerne *ne* tyrie
'No one irritate other person'
(WHom 10c:93)

If an NI subject occurred clause-initially and preceded a finite verb, the negative particle *ne* 'not' could be elided, as in (7), or it remained in the sentence, as in (5).

(7) **nænig anweald deaþes** him sceðþað
'no power of death hurts it'
(Bede 2 1.94.15)

If an NI subject followed a finite verb, on the other hand, the negative particle *ne* 'not' could not be omitted (see also Ingham, 2006). This contrast is true of NI objects. The distribution is similar to that of NPIs in Present-day English: NPIs cannot be used in subject position, or before the negative particle *not*, but they can be in object position, or after *not*, in negative sentences (see Blanchette, 2016). **Anybody didn't eat* and *John didn't eat anything* (Blanchette, 2016, 41-42). From the syntactic point of view, NIs can function more like NPIs if they are in the scope of the negative particle *ne* or *not*. In contrast, if they are out of the negative scope, they can be negative quantifiers (Blanchette, 2015 for Present-day English).

We can summarize that NIs in OE are morphologically negative but semantically quantifiers in the Negative Concord construction, as shown in (5), they can be counted as NPIs, as in (2), and that they are both morphologically and semantically negative if they occur out of the scope of the negative particle *ne*, as shown in (7).

## *References*

Blanchette, F. K. (2015). *English negative concord, negative polarity, and double negation* (Publication No. 866) [Doctoral dissertation, City University of New York]. CUNY Academic Works.

Blanchette, F. K. (2016). Subject-object asymmetries in English sentences with two negatives. *University of Pennsylvania Working Papers in Linguistics*, *22*, 40-50.

Hogg, R. (2002). *An introduction to Old English*. Edinburgh: Edinburgh University Press.

Ingham, R. (2006). On two negative concord dialects in Early English. *Language Variation and Change*, *18*, 241-266.

Mitchell, B. (1985). *Old English syntax* (2 vols.). Oxford: Clarendon Press.

Pintzuk, S., & Taylor, A. (2006). The loss of OV order in the history of English. In A. van Kemenade & B. Los (Eds.), *The handbook of the history of English* (pp. 249-278). Oxford: Blackwell.

Ringe, D., & Taylor, A. (2014). *The development of Old English*. Oxford: Oxford University Press.

Taylor, A., Warner, A., Pintzuk, S., & Beths, F. (2003). The York-Toronto-Helsinki parsed corpus of Old English prose. University of York, York. [YCOE]

Wallage, P. W. (2017). *Negation in Early English: Grammatical and functional change*. Cambridge: Cambridge University Press.

**Zhang, Jian**

***Xi’an Jiaotong University***

**zhangjane@stu.xjtu.edu.cn**

**Li, Yingyu**

***Xi’an Jiaotong University***

**hiwendy@mail.xjtu.edu.cn**

Type of Contribution: Work-in-Progress Report

# Comparison of adversative conjunctions in translation of the directional parallel corpus

*Abstract*

The sentences of English and Chinese are usually distinguished with parataxis and hypotaxis. In terms of the researches combining parallel corpus with translation studies, the issue stressed on the adversative meaning are convergent that few one-to-one mappings can be established across languages in both semantic and syntactic level (Granger & Altenberg, 2002; Dupont & Zufferey, 2017). The volatile nature of adversative conjunctions mainly referred that these conjunctions are frequently added or removed in the translated texts and sometimes make semantic or syntactic shifts in E-C translation (Altenberg, 2007; Li et al., 2017). Therefore, the transfer of meaning and grammar has the potential to explore. The main purpose is to study the transformation of adversative conjunctions in the E-C translation in semantic and syntactic levels between written and spoken texts, on the one hand, the features of adversative transformation can be explored; on the other hand, the similarities and dissimilarities could reveal whether text types would affect the translation of adversatives or not.

Thus, the research will be focused on four English conjunctions: *however, nevertheless*, *nonetheless* and *but* as synonyms. The research questions as follows:

(1) What are the most frequent correspondences of conjunctions in the E-C translation in terms of the relative frequency?

(2) As synonyms, which conjunctions are translated implicitly more?

(3) Which corpus transferred more in the translation of adversatives in terms of meaning and grammar?

Both two corpora are distracted from CQPweb BFSU and tagged for part of speech and aligned at the sentence level. In particular, the written material is Babel English parallel corpus with 327 English articles and their translations in Mandarin Chinese (Total English words 244,696; word types 22,273; total Mandarin words 275,361; word types 21,376). And the spoken material is from TED speeches parallel corpus where the transcriptions of oral presentations in English and their translations in Mandarin Chinese (Total English words 2,842,468; word types 64,280; total Mandarin words 2,707,571; word types 48,836). Although the word type and token in Babel are distinguished from that in TED, data can be standardized by statistical methods.

The results are as follows: above all, the relative frequency per million of each conjunction in both TED speeches and Babel parallel corpus would be presented. Next, the most frequent translated words of four conjunctions are presented.

| | **TED** | **Babel** |
|---|---|---|
| however | 62 | 417 |
| nevertheless | 11 | 37 |
| nonetheless | 11 | 25 |
| but | 5014 | 4367 |

Table 1: Relative frequencies per million words of the four Conjunctions in TED and Babel.

Table 1 shows the relative frequency per million of four conjunctions in English. It is worthwhile to mention that of these conjunctions, *but* is more prototypically used than the others in both TED and Babel. As for other conjunctions, their overall number in TED is lower than that in Babel. This distinction also reveals the different communication strategies between speakers and writers.

Table 2 shows the most frequent translated words of four conjunctions. The conjunction *但(是)* is frequently used as the correspondence of *but* which accounts for most in the originals in both TED and Babel corpus. It also suggests that the E-C translation share prototypical conjunction beyond the spoken and written type.

| | **however** | | **nevertheless** | | **nonetheless** | | **but** | |
|---|---|---|---|---|---|---|---|---|
| Most frequent translated words (%) | TED | Babel | TED | Babel | TED | Babel | TED | Babel |
| 但(是) | 35 | 21 | 26 | 33 | 34 | 17 | 60 | 64 |
| 然而 | 25 | 34 | 0 | 11 | 14 | 33 | 4 | 7 |
| 可(是) | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 3 |
| 不过 | 7 | 25 | 6 | 12 | 0 | 17 | 1 | 4 |
| Zero correspondence | 17 | 7 | 16 | 0 | 11 | 1 | 26 | 10 |
| Others | 12 | 9 | 35 | 44 | 41 | 17 | 9 | 12 |

Table 2: Most frequent correspondences among four conjunctions.

Apart from that, some mutual correspondences tend to be in low scores, such as *nevertheless* and *nonetheless*. Previous researches remind the result is due to the greater difference across language the lower score it represents (Altenberg, 1999). On the other hand, the meaning of adversatives is to some extent indispensable, in other words, the sentence meanings are easy to infer despite the absence of the conjunctions (Zufferey, 2016). Moreover, through a chi-square test of independence test, there are significantly difference in the zero correspondence of conjunctions *nevertheless*, *nonetheless* and *but* where TED is much more than that in Babel. This fact is that different types require the different degree of textual fidelity, since the conjunctions used less in TED than in Babel where the cohesion and connection are more essential in the written text (Lefer & Grabar, 2015). For example:

EN: Bio- and cybertechnologies are environmentally benign in that they offer marvelous prospects, while, *nonetheless*, reducing pressure on energy and resources (TED).

CN: 生物和网络技术为更好的改变环境提供了美妙的前景，得以减轻我们在能源方面的压力。

EN: *Nonetheless* , it's too early to count our such a bout(Babel).

CN: *无论如何*，现在就断定比赛的胜负还为时尚早。

The results will indicate that the conjunctions *nevertheless* and *nonetheless* have been implicated more than the others. Also, the adversatives in TED are transferred more than those in Babel. Moreover, it could be a confounding variable that the hybridity of text modes i.e. TED speeches are spoken, but the speech largely relies on a written script previously, which limits the translations of conjunctions.

## *References*

Altenberg, B. (1984). Causal linking in spoken and written English. *Studia Linguistica, 38*(1), 20-69.

Altenberg, B. (1999). Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: Studies in Honour of Stig Johansson* (pp. 249-268). Amsterdam: Rodopi.

Altenberg, B. (2007). The correspondence of resultive connectors in English and Swedish. *Nordic Journal of English Studies, 6*(1), 1-26.

Dupont, M., & Zufferey, S. (2017). Methodological issues in the use of directional parallel corpora. *International Journal of Corpus Linguistics, 22*(2), 270-297.

Granger, S., & Altenberg, B. (2002). *Lexis in contrast: Corpus-based approaches*. Amsterdam: John Benjamins.

Lefer, M.-A., & Grabar, N. (2015). Super-creative and over-bureaucratic: A cross-genre corpus-based study on the use and translation of evaluative prefixation in TED talks and EU parliamentary debates. *Across Languages and Cultures, 16*(2), 187-208.

Zufferey, S. (2016). Discourse connectives across languages. Factors influencing their explicit or implicit translation. *Languages in Contrast, 16*(2), 264-279.

李红英, 肖明慧, & 王永祥. (2017). 基于平行语料库的汉译英中连词显化的研究 a parallel corpus-based study of connective explicitation. 外国语文(四川外语学院学报), 033(006), 116-123.

**Zimmermann, Richard**

***University of Manchester***

**Richard.zimmermann@manchester.ac.uk**

Type of Contribution: Full Paper

## Word growth dispersion

A single corpus part measure of lexical dispersion

### *Abstract*

This paper introduces a measure of lexical dispersion across corpora, called *Dispersion of Word Growth* ($D_{WG}$). The following summarizes the essence of the calculation of the measure. First, determine the *Word Growth* (*WG*), an ordered set of the positions of all occurrences ($n$=the number of occurrences of the word) in a corpus ($N$=the size of the corpus).

(1) $WG$ = {position(1), position(2), …(position($n$)}

Second, get the distances between every occurrence starting from the second position.

(2) Distances = position($i$+1) - position($i$), $1 < i \leq n$

Third, calculate the deviations of the distances ($DD$) from a perfectly evenly distributed word, for which all occurrences would be equally spaced out at distance $N/n$.

(3) $DD$ = Distances - $N/n$

Finally, put the sum of the absolute $DD$s in relation to the sum of deviations that would be obtained from the worst possible distribution of the word in the corpus ($DD_{worst}$). The worst possible distribution arises if all word occurrences were placed immediately adjacent to each other.[1]

(4) $D_{WG}$ = sum($DD$) / sum($DD_{worst}$)

[1] The actual calculation of DWG involves several additional steps and details not specified here.

If all word occurrences follow a perfectly even distribution, the sum of their distance deviations will be 0, and 0 divided by the worst distribution is 0. If a word distribution actually follows the worst distribution, then the worst distribution divided by the worst distribution will be 1. Hence, $D_{WG}$ ranges between 0 and 1, with 0 indicating perfectly even dispersion and 1 maximally uneven distribution.

There is good reason to believe that $D_{WG}$ does in fact map onto the concept of lexical dispersion. Applications of $D_{WG}$ to natural language corpora return results that replicate findings from other dispersion measures. High frequency function words are well dispersed (*the, of*). High frequency items that are relatively badly dispersed are pronouns (*she, I*). Badly dispersed words are technical terms or names (*thyroid, Win*ston). Well dispersed words are those that straddle functional and lexical items (probably, make). However, the vast majority of words fall in between the latter two classes and show unpredictable dispersion behaviour (water, white, know). It is predominantly for those words that a dispersion measure is informative and relevant. Figure 1 illustrates the relation between DWG and frequency in the Brown corpus (Kučera & Francis, 1964).

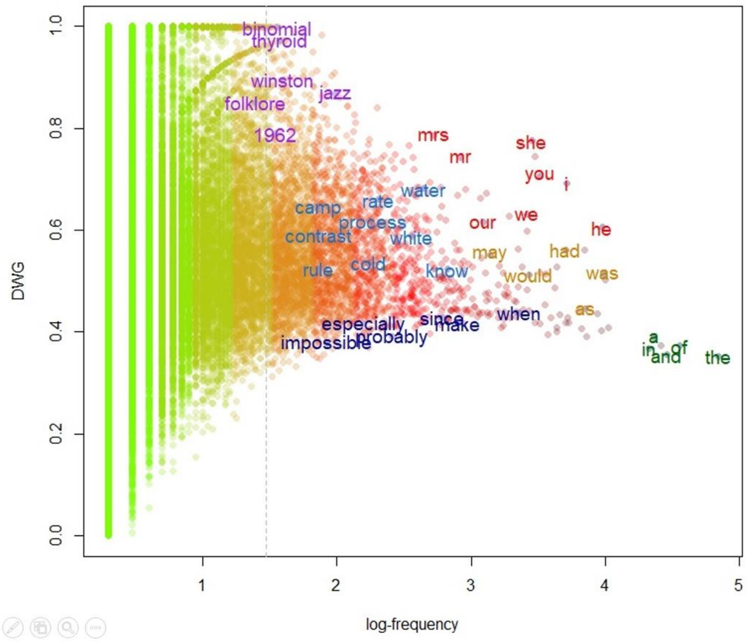


Figure 1: The distribution of DWG in Brown by word frequency (log scale).

Furthermore, $D_{WG}$ is highly correlated with other dispersion measures. For example, it shows a high correlation of $r$=0.78 with *DP* (Griess, 2008) for the Brown corpus.

Based on investigations of several corpora, a five-class scale of $D_{WG}$ seems reasonable. Values close to 0 indicating near-identical distances between occurrences do not normally exist in natural language. Experiments with artificial corpora show that randomly distributed words have a $D_{WG}$ of c. 0.37, which constitutes a benchmark for highly uniform dispersion. Extremely skewed word growth distributions, however, do exist in natural language. These are cases in which a word's occurrences are placed in close proximity to each other so that its distribution actually approaches its worst possible distribution. For example, the word *tent* in Brown shows extremely small distances between its occurrences, as shown in (5), which results in a large $D_{WG}$ value.

(5) There were umbrella tents, wall tents, cottage tents, station wagon tents, pup tents, Pop tents, Baker tents, tents with exterior frames (Brown, E-SkillsTradeHobbies, Sample E31 from Sports Age, 24:9 (1961))

The proposal is as follows: A $D_{WG}$ value of less than 0.45 may indicate very even dispersion, 0.45 to 0.525 even dispersion, 0.525-0.6 average dispersion, 0.6 to 0.7 uneven dispersion and a value greater than 0.7 very uneven dispersion.

Following the literature on comparisons between dispersion measures (e.g., Burch et al., 2017), the performance of $D_{WG}$ should be assessed for different criteria. Perhaps most importantly, the measure will be sensitive to the arrangement of the individual texts that make up a corpus. However, preliminary experiments on assessing the importance of this effect (with 15 corpus parts) suggest that variability caused by different corpus orders may be negligible, at least for large *n*. Next, dispersion measures of a word vary across different corpora. A preliminary experiment (with 10 natural 1m word corpora) shows that consistency is comparable for *DP* and $D_{WG}$. Finally, dispersion should be applicable for the correction of word frequencies, for instance, for the creation of learner vocabulary lists. Subtracting the $D_{WG}$ percentage from a word type's frequency results in a high correlation with adjusted frequency.

$D_{WG}$ has several additional advantages. The most important ones are the fact that it can be calculate easily from a single corpus part, relatively straightforward interpretability, scale invariance with respect to *n* and *N* for randomly distributed words, and a low correlation with frequency so that dispersion and frequency are separated into two distinct dimensions.

## *References*

Burch, B., Egbert, J., & Biber, D. (2017). Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science, 3*(2), 189-216.

Gries S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403-437.

Kučera, H., & Francis, N. W. (1964). *A standard corpus of present-day edited American English, for use with digital computers (Brown)*. Providence, RI: Brown University.

# ❯ ICAME 41 - Abstracts

## for the pre-conference workshop '*Crossing the borders: Complex contrastive data and the next generation'*

Workshop ran by Anna Čermáková (University of Birmingham), Signe Oksefjell Ebeling (University of Oslo), Magnus Levin (Linnaeus University), and Jenny Ström Herold (Linnaeus University)

*Wer fremde Sprachen nicht kennt, weiss nichts von seiner eigenen.*
(Those who do not know foreign languages, know nothing about their own.)
Johann Wolfgang von Goethe

Through the lens of other languages, we enrich our understanding of the complexities of English. Because of this, contrastive corpus-based studies have become an established research strand within the field of English corpus linguistics. As corpus research itself is moving beyond the space it has traditionally occupied, contrastive studies have reached a point where new materials and approaches are needed to explore new frontiers. Reflecting the overall conference theme, the ICAME 2020 contrastive workshop in Heidelberg is dedicated to new explorations and applications of complex multilingual data where English is contrasted with at least one other language. Previous contrastive workshops at ICAME have successfully investigated parallel and comparable corpora, largely focusing on lexicogrammatical features. Moreover, the papers at these workshops have mainly been corpus-based and synchronic, comparing English to only one other language (see, e.g., Ebeling & Hasselgård (eds) 2018, Janebová, Lapshinova-Koltunski & Martinková (eds) 2017, Egan & Dirdal (eds) 2017). This workshop therefore invited papers expanding the horizons of contrastive studies. This involved new types of synchronic or diachronic corpus data, new language pairs - in particular going beyond the traditional two-language perspective -, new areas of investigation such as semantics, pragmatics and phraseology combined with more corpus-driven methods and interdisciplinary approaches.

### *References*

Ebeling, S.O., & Hasselgård, H. (Eds). (2018). Corpora et Comparatio Linguarum: Textual and Contextual Perspectives. Bergen Language and Linguistics Studies (BeLLS) Vol 9(1). (ICAME 38, Prague 2017).

Egan, T., & Dirdal, H. (Eds). (2017). Cross-linguistic correspondences. Amsterdam: John Benjamins. (ICAME 36, Trier 2015).

Janebová, M., Lapshinova-Koltunski, E., & Martinková, M. (Eds). (2017). Contrasting English and other Languages through Corpora. Newcastle: Cambridge Scholars Publishing. (ICAME 37, Hong Kong 2016).

**Axelsson, Karin**

***University of Skövde***

**karin.axelsson@his.se**

Type of Contribution: Pre-Conference Workshop Submission

## Semicolons in English, Swedish and Norwegian fiction texts

Rules, recommendations and real usage

*Abstract*

Contrastive linguistics has so far mainly dealt with lexico-grammatical features of written texts (e.g. Egan & Dirdal, 2017), whereas much less attention has been paid to punctuation. However, there has been a recent renewed general interest in punctuation, demonstrated through the first conference solely devoted to punctuation, held in Regensburg in May 2019.

Punctuation is often regarded as marginal in grammars: the major grammars of Swedish (Teleman et al., 1999) and Norwegian (Faarlund et al., 1997) as well as Biber et al. (1999) hardly mention it, and Quirk et al. (1985) relegate it to an appendix. Huddleston and Pullum (2002) devote a whole chapter to it, but declare that they have given punctuation special treatment: "we [...] have given greater weight to the *prescriptions* of major style manuals than we have in the chapters on *grammar*" (2002, p. 1727, my italics). Such prescriptive attitudes to punctuation may explain why there are so few descriptive corpus-based studies on punctuation, both in general and within contrastive linguistics.

The aim of this study is to compare the use of semicolons in original fiction texts in English, Swedish and Norwegian, in relation to guidebook rules; translations are used for assistance in the analysis. The data comes from the *English-Swedish Parallel Corpus* (Aijmer et al., 2001) and the *English-Norwegian Parallel Corpus* (Johansson et al., 1999/2002).

The rules for the semicolon are similar in the three languages, but the frequencies differ (per 10,000 words: 24 for English, 8 for Norwegian and 6 for

Swedish). The guidebooks mention two reasons for using a semicolon: connecting two independent clauses, as in (1), and separating items in a series, in particular when the items themselves contain commas, as in (2):

(1) There was no doubt; the diamonds were gone. (ESPC/ENPC: EO FF1)

(2) Essential Items: Furniture; Fittings, suitable for two-bedroomed council house and pensioner's bungalow. (ESPC/ENPC: EO ST1)

The first type is predominant in all three languages. The semicolon is then an alternative to the full stop; it is up to writers to decide whether to use it or not. However, the general attitude to semicolons is more positive in English; it is, for example, recommended between certain independent clauses (*Chicago Manual of Style*, 2017). For Swedish, it is said to be a matter of taste (Språkrådet, 2017), and for Norwegian, a resource for variation (Søyland & Fretland, 2015). The second function is used to some extent in English but very seldom in Swedish and Norwegian fiction texts.

The data reveals that there is a large third group of semicolon use: between a fragment/ellipted clause and an independent clause, or between two fragments/ ellipted clauses, as in (3). For semicolons of type 1, such cases account for 38% in the Norwegian texts and 28% in the Swedish texts but only 19% in the English texts.

(3) Everything undefined; except the eyes. (ESPC/ENPC: EO NG1)

This study investigates both the syntactic types of semicolon use and the semantic relations between the independent sentences/fragments before and after the semicolon.

## *References*

Aijmer, K., Altenberg, B., & Svensson, M. (2001). *English-Swedish parallel corpus: Manual.* Retrieved from https://sprak.gu.se/english/research/researchactivities/corpuslinguistics/corpora-at-the-dll/espc.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

*Chicago Manual of Style*. (2017). (17th ed.). The University of Chicago Press.

Egan, T., & Dirdal, H. (Eds.). (2017). *Cross-linguistic correspondences*. Amsterdam: John Benjamins.

Faarlund, J. T., Lie, S., & Vannebo, K. I. (1997). *Norsk referansegrammatikk* [Norwegian reference grammar]. Universitetsforlaget.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Johansson, S., Ebeling, J., & Oksefjell, S. (1999/2002). *English-Norwegian Parallel Corpus: Manual.* Retrieved from https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/enpc/ENPCmanual.pdf.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Harlow: Longman.

Søyland, A., & Fretland, J. O. (2015). *Norske skriveregler: reglene du trenger for å skrive på papir og skjerm* [Norwegian writing rules: the rules you need to write on paper and on screen] (D. Ellingsen, Trans.). Samlaget.

Språkrådet. (2017). *Svenska skrivregler* [Swedish Writing Rules] (O. Karlsson Ed., 4 ed.). Liber.

Teleman, U., Hellberg, S., & Andersson, E. (1999). *Svenska Akademiens grammatik* [The grammar of the Swedish Academy]. Svenska Akademien.

**Ebeling, Signe Oksefjell**

***University of Oslo***

**s.o.ebeling@ilos.uio.no**

Type of Contribution: Pre-Conference Workshop Submission

## Hope for the future

### An analysis of HOPE/HÅP(E) across genres and languages

*Abstract*

In a previous contrastive study of English and Norwegian football match reports it was found that the cognates *hope* and *håp* 'hope' featured as keywords in texts reporting on defeat (Ebeling, 2019). The reason for this frequent use of *hope*/*håp* in the defeat section of the English-Norwegian Match Report Corpus (ENMaRC) was attributed to the items' use in contexts where hopes are dashed, as in examples (1) and (2).

(1) However those *hopes* were dashed on 55 minutes when the Gunners added a second. (CPFC)

(2) Det tente et ørlite *håp* som ble knust desto mer brutalt fem minutter etter. (VFK)
Lit.: That lit a tiny hope that was dashed even more brutally five minutes later

Drawing on data from the ENMaRC and the English-Norwegian Parallel Corpus+ (ENPC+) of fiction texts, this study seeks to dig deeper into the lemmas HOPE and HÅP(E) - both nouns and verbs - by contrasting their conditions of use across languages (English and Norwegian) and genres (match reports and fiction). More specifically, the study seeks answers to the following questions:

- To what extent are the lemmas used similarly in English and Norwegian?
- To what extent are the lemmas used similarly in match reports and fiction?

- To what extent does the use of different types of “contrastive” corpora contribute to our cross-linguistic knowledge of the lemmas?

The study starts with an overview of the rather complex data under investigation. The size of the corpora differ somewhat: the ENMaRC contains around 990,000 words from the English Premier League and 315,000 words from the Norwegian ‘Eliteserie’, while the ENPC+ is more balanced with around 1.3 million words of English and Norwegian original texts with their translations. With both a comparable corpus (ENMaRC) and a bidirectional parallel corpus (ENPC+) at hand a sound contrastive approach is ensured. Further, a Mutual Correspondence (Altenberg, 1999) of a staggering 93.3% in the ENPC+ demonstrates that the lemmas are very good cross-linguistic matches of each other and they can safely serve as the starting point of a contrastive analysis.

The distribution of noun and verb uses is more similar across the two languages than across the two genres. In both English and Norwegian fiction, the verb use outnumbers the noun use by 28 and 20 occ. per 100,000 words vs. 6 and 8 occ., respectively. In the match reports we see the opposite trend, with the noun use outnumbering the verb use, particularly in English: 23 vs. 9 occ. per 100,000 words and, in Norwegian, 20 vs. 15.

Following this general overview of the data, a qualitative analysis of the actual uses of the lemmas will be carried out. Preliminary scrutiny of concordance lines suggests that the lemmas have the potential to feature in a number of different (phraseological) contexts in both languages. This is particularly evident in the case of the nouns: HOPE and HÅP are confirmed to be relatively more frequent in the match reports than in the fiction corpus in contexts such as (1) and (2).

### *References*

Altenberg, B. (1999). Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 249-268). Atlanta: Rodopi.

Ebeling, S. O. (2019). The language of football match reports in a contrastive perspective. In M. Callies & M. Levin (Eds.), *Corpus approaches to the* *language of sports: Text, media, modalities* (pp. 37-62). London: Bloomsbury Academic.

**Hasselgård, Hilde**

***University of Oslo***

**hilde.hasselgard@ilos.uio.no**

Type of Contribution: Pre-Conference Workshop Submission

## Lexicogrammar through colligation

Noun + preposition in English and Norwegian

*Abstract*

This study uses sequences of PoS tags as a window into cross-linguistic syntactic differences. The selected tag sequence is noun plus preposition, most frequently realizing either noun with postmodifying PP or chance sequences of noun + PP with adverbial function, illustrated by (1) and (2) from the English-Norwegian Parallel Corpus (ENPC) on which this study is based. Both examples have congruent translations, indicating potential similarity between English and Norwegian PPs.

(1) Nor did he enjoy his *meetings with* Dr Forestier… (BC1)
Han likte heller ikke *konsultasjonene hos* doktor Forestier …

(2) Og han hadde lagt *klærne på* en stein mye lenger opp. (HW2)
And he had laid his *clothes on* a rock much nearer the grove.

The research questions are as follows:

- What are the syntactic functions of PPs following a noun in Norwegian and English?
- What meanings do the PPs convey?
- Are there quantitative and qualitative differences between the languages?
- To what extent are translations congruent?

English is expected to have more postmodifying PPs and Norwegian to have more adverbial PPs. This is based on a study of postmodifying *of*-phrases and their (frequently non-congruent) Norwegian correspondences (Hasselgård, 2016). Furthermore, the claim that English is more nominal while Norwegian is

more verbal/sentential (e.g. Nordrum, 2007; Behrens, 2014), might promote postmodifying PPs in English and clause-level adverbials in Norwegian.

A novelty of this study is its use of tag sequences as a starting point for the investigation, which has not been common in cross-linguistic corpus studies (though see Wilhelmsen, 2019 and monolingual studies of L1 and L2 performance, e.g. Granger & Rayson, 1998; Granger & Bestgen, 2014).

Considering the high numbers of noun+preposition sequences in the ENPC and the need for manual post-processing of the data, random samples of 500 from either language were extracted through the Glossa search interface. As expected, the most important functions of the PPs were postmodifier and adverbial. The languages differ: in English the number of postmodifiers is about twice that of adverbials, but in Norwegian the number of adverbials is very close to that of postmodifiers. Other, infrequent, syntactic functions include complex prepositions and multi-word verbs not recognized by the tagger as such (e.g. *in front of*, *legge merke til* 'make note of'). In addition there were some tagging errors concerning both nouns and prepositions.

Locative meanings are more frequent in Norwegian than in English in both adverbial and postmodifiying PPs. English has more PPs modifying support nouns (Sinclair, 1991) and PPs in possessive constructions. Larger proportions of the English adverbials have temporal and manner meanings. Other meanings are more similarly distributed.

The degree of congruence in translation varies across translation directions and syntactic functions. Adverbials are translated congruently in 75-80% of the cases in both directions. Translations of postmodifiers are congruent more often from Norwegian to English (c. 75%) than from English to Norwegian (c. 55%). This indicates greater cross-linguistic differences in postmodifying than adverbial PPs.

## *References*

Behrens, B. (2014). Nominalization: A case study of linguistic text conventions in comparable and parallel texts- English and Norwegian. In S.O. Ebeling, A. Grønn, K.R. Hauge & D. Santos (Eds.), *Corpus-based*

*Studies in Contrastive Linguistics, Oslo Studies in Language 6*(1), 143-160.

English-Norwegian Parallel Corpus (ENPC), fiction. Retrieved from: http://www.hf.uio.no/ilos/english/services/omc/.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. I*nternational Review of Applied Linguistics in Language Teaching, 52*(3), 229-252.

Granger, S., & Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (Ed.), *Learner English on computer* (pp. 119-131). London: Longman.

Hasselgård, H. (2016). *The way of the world*: The colligational framework 'the N1 of the N2' and its Norwegian correspondences. *Nordic Journal of English Studies, 15*(3), 55-79.

Johansson, S. (2007). *Seeing through multilingual corpora*. Amsterdam: Benjamins.

Nordrum, L. (2007). *English lexical nominalizations in a Norwegian-Swedish contrastive perspective*. PhD thesis, University of Göteborg.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik J. (1985). *A comprehensive grammar of the English language.* London: Longman.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Wilhelmsen, A. (2019). Pretty complete or completely pretty? Investigating degree modifiers in English and Norwegian original and translated text. MA thesis. Faculté de philosophie, arts et lettres, Université catholique de Louvain, 2019.

**Johansen, Stine Hulleberg**

***University of Oslo***

**s.h.johansen@ilos.uio.no**

Type of Contribution: Pre-Conference Workshop Submission

## A contrastive study of speaker- and hearer-oriented hedging strategies in Norwegian and English conversations

### *Abstract*

Hedging strategies are discourse strategies that reduce the force or truth of an utterance (Kaltenböck, Mihatsch, & Schneider, 2010, p. 1) and are a complex phenomenon, which requires a high degree of sophistication to master, even in one's native language (Markkanen & Schröder, 1997, p. 13). Their complexity and versatility make them particularly interesting to study across languages, but also represent a challenge for traditional corpus approaches, due to the frequent mismatch between form and function.

The current paper is part of a contrastive study of hedging strategies in spoken Norwegian and English informal conversations based on four corpora (the BigBrother Corpus, the Norwegian Speech Corpus, the Nordic Dialect Corpus and the British National Corpus, 2014). In the larger study, 1,439 hedging strategies were retrieved from the corpora using a probe, i.e. an element, such as a word, phrase or tag, used to find other elements which cannot easily be found in a corpus (Hunston, 2002, p. 62). After classification, the results showed that 15.7% of the Norwegian strategies were speaker-oriented, i.e. reducing the commitment of the speaker towards the utterance (Hyland, 1996, p. 443), as in (1), and 28.8% were hearer-oriented, i.e. seeking common ground between the interlocutors (Hyland, 1996, p. 446) as in (2). The English results pointed in the opposite direction: 25.9% were speaker-oriented and 12.6% were hearer-oriented.

(1) […] I *just* find it *a bit* draining *like* (BNC2014 274)

(2) […] Jeg er *jo* ikke *så* intelligent jeg (NDC 568)

[…] I am *[particle]* not *so* intelligent I

These differences form the backdrop of the current study, which aims to discover:

(1) How do English and Norwegian differ in regard to speaker-oriented and hearer-oriented hedging strategies in informal conversations?

(2) What are the potential implications of such differences?

All speaker- and hearer-oriented strategies were categorized based on their form. 73.5% of the English speaker-oriented strategies were some kind of pragmatic particle, mainly *like* and *just*. *Just* and *like* may play a face-saving role and downplay the delivery of a message (Beeching, 2016, p. 76, p. 132). In the Norwegian material, about 50% were pragmatic particles, mainly *bare* ('just') and *liksom* ('like'), which share many of the same features as *just* and *like* (Hasund, Opsahl, & Svennevig, 2012, p. 43; Hasund, 2003, p. 123).

The greatest cross-linguistic difference was found in the hearer-oriented category. 77.9% of the Norwegian strategies were pragmatic particles, mainly *jo* ('after all', 'of course'). In English, however, there were no pragmatic particles, but a high proportion of first/second personal pronoun + cognitive verb, e.g. *you know*. *Jo* does not have a direct counterpart in English, but serves some of the same pragmatic functions as *you know* (Berthelin & Borthen, 2019, p. 8). Such differences may cause problems in second-language communication as "modal meaning tends to disappear in the translation process" (Aijmer, 1996, p. 395). The most frequent expressions in the English material were compared in an intervarietal study of English by Norwegian learners with data from the LINDSEI-no and LOCNEC corpora. The results showed that expressions like *you know* were underused by Norwegian learners.

## *References*

Aijmer, K. (1996). Swedish modal particle in a contrastive perspective. *Language Sciences, 18*(1-2), 393-427.

Beeching, K. (2016). *Pragmatic markers in British English: Meaning in social interaction.* Cambridge: Cambridge University Press.

Berthelin, S. R., & Borthen, K. (2019). The semantics and pragmatics of Norwegian sentence-internal *jo*. *Nordic Journal of Linguistics, 42*(1), 3-30.

BigBrother-korpuset. Tekstlaboratoriet, ILN, Universitetet i Oslo.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL),* 225-246.

Gilquin, G., De Cock, S., & Granger, S. (2010). Louvain International Database of Spoken English Interlanguage (CD-ROM + handbook). Louvain-la-Neuve: Presses universitaires de Louvain.

Hasund, I. K. (2003). The discourse markers *like* in English and *liksom* in Norwegian teenage language: a corpus-based, cross-linguistic study. Unpublished doctoral dissertation. University of Bergen/Agder University College.

Hasund, I. K., Opsahl, T., & Svennevig, J. (2012). By three means: The pragmatic functions of three Norwegian quotatives. In I. Buchstaller & I. van Alphen (Eds.), *Quotatives: Cross-linguistic and cross-disciplinary perspectives* (37-68). Amsterdam: John Benjamins.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics, 17*(4), 433-454.

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A., & Vangsnes, Ø. A. (2009). The Nordic Dialect Corpus: An advanced research tool. In K. Jokinen & E. Bick (Eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*.

Kaltenböck, G., Mihatsch, W., & Schneider, S. (2010). *New approaches to hedging.* UK: Emerald Bingley.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics, 22*(3), 319-344.

Markkanen, R., & Schröder, H. (1997). *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts.* Berlin: Walter de Gruyter.

Norsk talespråkskorpus - Oslodelen. In Tekstlaboratoriet, ILN, Universitetet i Oslo.

**Krielke, Marie-Pauline**

***Saarland University***

***mariepauline.krielke@uni-saarland.de***

Type of Contribution: Pre-Conference Workshop Submission

# Close relatives?

A diachronic study of relativizers in 250 years of scientific English and German

*Abstract*

This paper presents a contrastive diachronic study on relativizer use in English and German scientific writing (SciEng vs SciGer) between 1650 and 1900. This time period is particularly interesting, since academic disciplines and with it scientific discourse started to develop (Görlach, 2008). English and German are closely related languages, but they have largely different trajectories of scientific discourse. SciEng became institutionalized with the foundation of the Royal Society (1660), while a comparable institution for scientific publications in German did not exist at the time.

Previous research has shown that SciEng developed towards denser ways of encoding, i.e. less verbal, more nominal style (Biber, 2006; Aarts et al., 2012; Biber & Gray, 2016; Degaetano-Ortlieb et al., 2019). SciGer was still under a strong Latin influence pushing the development of unambiguous conjunctions and leading to a complex hypotactic syntax with multiple embeddings until 1800. A trend of detangling this intricate syntax is observed afterwards (Admoni, 1967; Möslein, 1974; Beneš, 1981; Habermann, 2001). To trace this development, we take relativizers as an example, since they are indicators of more explicit style (expressing in a clause what could often be said in an NP).

Our analyses of SciEng are based on the Royal Society Corpus (RSC V4.0; Kermes et al., 2016) with approx. 32 million tokens from 1665 to 1869. For German we use a scientific sub-corpus of Deutsches Textarchiv (DTA,

Geyken et al., 2018) between 1650 and 1900, spanning approx. 80 million tokens. We look at relative frequencies and diversity of the paradigm of relativizers. To measure diversity, we use entropy, which “increases with a higher number of members of the paradigm, as well as with greater similarity of the probabilities of the members” (Milin et al., 2009, p. 6). For SciEng the relativizer paradigm should shrink, while SciGer should expand the paradigm towards the end of 18th century and shrink afterwards. In terms of overall relative frequencies, we expect SciEng to reduce relativizer use, while SciGer relativizers should show an increase towards the late 18th century and a decrease afterwards.

Our analyses show that SciEng reduces relativizer diversity over time for the benefit of entropy reduction (62% drop from 1650 to 1850), while SciGer shows an overall much higher entropy, relatively stable between 1650 and 1800 and dropping afterwards. SciEng decreases overall relativizer use pointing at a denser style, while SciGer keeps and intensifies use (peak in 1750 and drop afterwards) pointing at a more explicit, verbal style. Further, we measure average number of relativizers per sentence. While for SciEng relativizers per sentence continuously decrease over time, SciGer shows a peak of embeddedness in 1700. Correlation shows that lower grade of embeddedness correlates with lower frequency overall in SciEng (cor: 0.998, p-value = 0.0001) but not in SciGer (cor: 0.430, p-value = 0.3943). We conclude that SciEng undergoes a paradigm reduction paired with an overall densification on the syntactic level. SciGer shows an extended period of diversification and intensification in the use of the relativizer paradigm (until 1800) reflecting the syntactic intricacy typical for the time and a clear decrease in diversity and frequency towards the end of the 19th century.

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 232722074 - SFB 1102 / Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 232722074 - SFB 1102

## *References*

Aarts, B., López-Couso, M. J., & Méndez-Naya, B. (2012). Late Modern English syntax. In A. Bergs & L. J. Brinton (Eds.), *Historical linguistics of English:*

*An international handbook* (pp. 869-887). Vol. 1. Berlin: Mouton de Gruyter.

Admoni, V. G. (1990). *Historische Syntax des Deutschen*. Tübingen: Niemeyer.

Beneš, E. (1981). Die formale Struktur der wissenschaftlichen Fachsprachen in syntaktischer Hinsicht. In T. Bungarten (Ed.), *Wissenschaftssprache. Beiträge zur Methodologie, theoretischen Fundierung und Deskription* (pp. 185-211). München: Fink.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins Publishing.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge: Cambridge University Press.

Degaetano-Ortlieb, S., Kermes, H., Khamis, A., & Teich, E. (2019). An information-theoretic approach to modeling diachronic change in scientific English. In C. Suhr, T. Nevalainen & I. Taavitsainen (Eds.), *From data to evidence in English language research, language and computers* (pp. 258-281). Leiden: Brill.

Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C., & Wiegand, F. (2018). Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In H. Lobin, R. Schneider & A. Witt (Eds.), *Digitale Infrastrukturen für die germanistische Forschung* (pp. 219-248). Berlin: Mouton de Gruyter.

Görlach, M. (2008). *Text types and the history of English*. Berlin/Boston: Mouton de Gruyter.

Habermann, M. (2011). *Deutsche Fachtexte der frühen Neuzeit. Naturkundlich-medizinische Wissensvermittlung im Spannungsfeld von Latein und Volkssprache*. Berlin/Boston: Mouton de Gruyter.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J., & Teich, E. (2016). The Royal Society Corpus: From uncharted data to corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). Portorož, Slovenia.

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. H. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins and J. Blevins

(Eds.), *Analogy in grammar: Form and acquisition* (214-253). Oxford: Oxford University Press.

Möslein, K. (1974). Einige Entwicklungstendenzen in der Syntax der wissenschaftlich-technischen Literatur seit dem Ende des 18. Jahrhunderts. *Beiträge zur Geschichte der deutschen Sprache und Literatur, 94*, 156-198.

**Lastres-López, Cristina**

***University of Santiago de Compostela***

**cristina.lastres@usc.es**

Type of Contribution: Pre-Conference Workshop Submission

## Conditionals in conversation

A contrastive corpus-based study in English, French and Spanish

*Abstract*

Conditionals in English have long attracted scholarly attention from many different perspectives. Despite this, there are relatively few studies aiming to compare conditional constructions in English with other languages. Exceptions to this are Dancygier (1985) and Polańska (2006), on English and Polish; Lavid (1998), on English, Italian and German; and Hasselgård (2014), on English and Norwegian. In addition, some studies that have adopted a contrastive perspective have focused on very specific registers that lie on the domain of languages for specific purposes (see, for example, Visconti, 1996, on English and Italian legal discourse; and Carter-Thomas, 2007, on English and French medical discourse). This paper, therefore, will contribute to fill a gap in the contrastive linguistics scenario, where, to the best of my knowledge, conditional constructions in English, French and Spanish have not been compared. The analysis centres on the prototypical markers of conditionality - *if* in English, and *si* in French and Spanish - on the grounds that prior research has demonstrated that these are the dominant in the three languages, with other conditional markers playing a much more marginal role (Lastres-López, 2019).

The framework proposed examines conditionals in the light of the metafunctions proposed in Systemic Functional Linguistics (Halliday & Matthiessen, 2014): ideational, interpersonal and textual; in order to encompass the discourse-pragmatic multifunctionality of the constructions under analysis. The aim of my paper is to elucidate the uses and functions of conditionals in the three aforementioned languages in conversation, where

these constructions are pragmatically rich and often depart from prototypical cause-consequence patterns, as in (1) to (3). The examples illustrated in (1) to (3) below serve different functions as interpersonal devices in discourse. For these constructions, I propose a twofold classification: (i) as to whether they convey stance or engagement in discourse (Hyland, 2005); and (ii) in terms of the subfunction they fulfil in interaction, ranging from politeness to metalinguistic comments, among others.

(1) And if I remember rightly, you had jaundice, didn't you? (ICE-GB: S1A-028 #051:1:A)

(2) C'est fixé sur un objet, si tu veux (C-ORAL-ROM, ffamcv08)
'It's fixed in an object if you want'

(3) Si no has comido, hijo mío, tengo potaje (C-ORAL-ROM, efamdl10)
'If you haven't eaten, my son, I have stew'

The methodology adopted is corpus-based. Data are extracted from the conversations subcorpus of the British component of the International Corpus of English (ICE-GB) (Nelson et al., 2002) and from the same subcorpora of the French and Spanish components of the Integrated Reference Corpora for Spoken Romance languages (C-ORAL-ROM) (Cresti & Moneglia, 2005), which are similar in terms of corpus structure and size. Preliminary results suggest a path of pragmaticalization for the constructions under analysis. In addition, corpus findings show variation in the use of conditionals across the three languages examined. English and Spanish show a preference for interpersonal conditionals in conversation as opposed to French, which uses these constructions for cause-consequence patterns more frequently.

### *References*

Carter-Thomas, S. (2007). The 'iffiness' of medical research articles. A comparison of English *if* and French *si*. In K. Fløttum (Ed.), *Language and discipline perspectives on academic discourse* (pp. 150-175). Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Cresti, E., & Moneglia, M. (Eds.). (2005). *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Language*s. Amsterdam: John Benjamins.

Dancygier, B. (1985). *If*, *unless*, and their Polish equivalents. *Papers and Studies in Contrastive Linguistics, 22*, 65-72.

Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's introduction to functional grammar*. London: Routledge.

Hasselgård, H. (2014). Conditional clauses in English and Norwegian. In H. P. Helland & C. Meklenborg Salvesen (Eds.), *Affaire(s) de grammaire. Mélanges offerts à Marianne Hobæk Haff à l'occasion de ses soixante-cinq ans* (pp. 185-204). Oslo: Novus Forlag.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, *7*(2), 173-192.

Lastres-López, C. (2019). Conditionals in spoken courtroom and parliamentary discourse in English, French and Spanish: A contrastive analysis. In T. Fanego & P. Rodríguez-Puente (Eds.), *Corpus-based research on variation in English legal discourse* (pp. 51-78). Amsterdam: John Benjamins.

Lavid, J. (1998). Discourse functions of conditionals in multilingual instructions: A corpus study on ordering variants. *Papers and Studies in Contrastive Linguistics, 34,* 285-301.

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: The British component of the International Corpus of English*. Amsterdam: John Benjamins.

Polańska, I. (2006). *Expressing condition in English and in Polish*. Krakow: Jagiellonian University Press.

Visconti, J. (1996). On English and Italian complex conditional connectives: Matching features and implicatures in defining semanto-pragmatic equivalence. *Language Sciences, 18*(1-2), 549-573.

**Levin, Magnus**

***Linnaeus University***

**magnus.levin@lnu.se**

**Ström Herold, Jenny**

***Linnaeus University***

**jenny.strom.herold@lnu.se**

Type of Contribution: Pre-Conference Workshop Submission

## On parentheticals from a multilingual perspective (or: how to be explicit within brackets)

*Abstract*

In this paper, we investigate the use of brackets in order to explore written conventions in English, German and Swedish. More specifically, we zoom in on the functions of brackets as (i) interpersonal or information-condensing devices (Bredel, 2018, p. 11) and (ii) explicitation strategies for translators (Baumgarten, Meyer & Özçetin, 2008). By drawing on new trilingual data subsuming both originals and translations, we can disentangle language-specific and translation-related features and preferences of this under-researched punctuation mark.

Our material is collected from the novel Linnaeus University English-German-Swedish corpus (LEGS) (Ström Herold & Levin, 2019), which contains popular non-fiction books from the 2000s. Different types of bracketed structures from the LEGS data are shown below, moving from condensed elaborations to more extensive reader-oriented comments.

(1) […] the Koch-backed organization Americans for Prosperity *(AFP)* launched a campaign […].

(2) […] the government wasted hundreds of millions *(at least)* trying to clean up the unnecessary messes.

(3) Andrew *(who bears a striking resemblance to Baldrick from Blackadder)* came up with the cunning plan of mass-marking the bees […].

Our results regarding originals indicate that brackets are the most frequent in English and the rarest in Swedish originals. English and Swedish brackets contain significantly more clausal, reader-oriented comments (as in (3)) than German (cf. House 2011), which instead favours shorter content-oriented phrases (as those seen in (1) and (2)). English brackets typically contain more words than German and Swedish ones. Two very different reasons can be identified for the frequent use of brackets: (formal) information condensation (cf. Biber & Gray, 2016, p. 205) in phrases and (informal) addressee-oriented commentary (House, 2011, p. 195) in clauses. Clauses are rarely bracketed in German texts, a finding reflecting the increasing German avoidance of verb-finite subordinate clauses (Bisiada, 2013).

Regarding translations, two main trends emerge: first, most brackets in originals are retained (c. 80%) in translations and second, most translations contain more brackets than originals. The high retention rate is in line with previous findings on punctuation in translation (Ström Herold, Levin & Tyrkkö, forthcoming; Wollin, 2018). The high rate of additions depends on translators often adding new information in the form of, e.g., name variants (*from Antwerp to Louvain* > *von Antwerpen bis Louvain (Löwen)*) or explanations of terms (*hemoglobinet* > *hemoglobin (a component of blood)*). As seen in these examples, added brackets are mostly short phrasal elaborations (Newmark, 1988, p. 92).

The study of brackets in trilingual translation and comparable data allows the extraction of language preferences and translation-induced changes. Brackets provide a window both on how writers explicitly become involved in their texts and what translators deem as requiring additional information.

## *References*

Baumgarten, N., Meyer, B., & Özçetin, D. (2008). Explicitness in translation and interpreting: A critical review and some empirical evidence (of an elusive concept). *Across Languages and Cultures, 9*(2), 177-203.

Biber, D., & Gray B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge: Cambridge University Press.

Bisiada, M. (2013). Changing conventions in German causal clause complexes: A diachronic corpus study of translated and non-translated business articles. *Languages in Contrast, 13*(1), 1-27.

Bredel, U. (2018). Das Interpunktionssystem des Deutschen. *Studia Neophilologica, 90*(1), 7-23.

House, J. (2011). Using translation and parallel text corpora to investigate the influence of Global English on textual norms in other languages. In A. Kruger, K. Wallmach & J. Munday (Eds.), *Corpus-based translation studies* (187-208). London: Bloomsbury.

Newmark, P. (1988). *A textbook of translation*. New York: Prentice Hall.

Ström H., Levin, J., & Levin, M. (2019). *The Obama presidency*, *the Macintosh keyboard* and *the Norway fiasco*: English proper noun modifiers in German and Swedish contrast. *English Language and Linguistics, 23*(4), 827-854.

Ström Hero, J., Levin, M., & Tyrkkö, J. (forthcoming). The colon in German, English and Swedish: A contrastive corpus-based study. Paper presented at Punctuation seen internationally, May 3-4, 2019, University of Regensburg.

Wollin, L. (2018). Punctuation: Providing the setting for translation? *Studia Neophilologica*, *90*(1), 37-49.

**Põldvere, Nele**

***Lund University***

**nele.poldvere@englund.lu.se**

**Johansson, Victoria**

***Lund University***

**victoria.johansson@ling.lu.se**

**Paradis, Carita**

***Lund University***

**carita.paradis@englund.lu.se**

Type of Contribution: Pre-Conference Workshop Submission

## On the importance of audio material in spoken linguistics

A case study of the London-Lund Corpus 2

*Abstract*

The London-Lund Corpus 2 (LLC-2) is a new corpus of spoken British English, modeled on the same principles as the world's first machine-readable spoken corpus, the London-Lund Corpus (LLC-1). An important novelty of LLC-2 is that its transcriptions are released together with the audio files. This feature is important because, in contrast to LLC-1, the transcriptions in LLC-2 are orthographic and not annotated for prosodic features such as the pitch contour. In addition, similar to many other well-known corpora of spoken English (e.g., the British National Corpus), the transcriptions in LLC-2 contain basic representations of spoken features such as pauses, overlaps, non-verbal vocalizations, to name a few. Thus, the release of the audio files alongside the transcriptions is an innovation that allows users to extend the transcriptions relative to their own research interests, whether these concern prosodic information or other aspects of speech production. This paper has two aims,

namely to (i) describe and discuss the main challenges of preparing the LLC-2 audio material for public release, and (ii) demonstrate the importance of the LLC-2 audio material by means of a case study of the prosodic and temporal aspects of impromptu speech.

The most important challenge encountered in the preparation of the LLC-2 audio material for public release was the anonymization of personal information (e.g., names, workplaces, addresses) in the audio files without losing important linguistic information such as prosody. However, the anonymization procedure was not straightforward because it required careful manipulation of the speech signal. Moreover, the approaches adopted in other spoken corpora have not been completely satisfactory; for example, the decision to mute personal information in the 1994 British National Corpus ensured a reliable result, but also led to the loss of prosodic information in the original speech signal. The solution in LLC-2 is based on a Praat script written and developed by Hirst (2013). The script replaces all personal information in the recordings with *hum* sounds, which make the information incomprehensible, while at the same time retaining the pitch contour and intensity level of the original. The resulting audio files are thus compatible with the ethical requirements of data protection and privacy, and they also retain linguistically useful information such as prosody and other aspects of speech production.

In order to demonstrate the importance of the public release of the LLC-2 audio material, we carried out a case study investigating a powerful source of coordination in language, namely dialogic resonance, or when speakers reproduce constructions from prior turns. Consider (1) where resonance is achieved through B's choice of words and structures. The square brackets in the example represent overlaps.

(1) A: she hasn't hitherto been particularly interested in religious things
[has she]
B: [you mean] she hasn't particularly been up at seven AM

According to Du Bois (2014), dialogic resonance draws on conscious strategies of interpersonal engagement. While Du Bois (2014) acknowledges the role of automatic priming, this is not tested in his work. Instead, priming is the central mechanism of Garrod's and Pickering's (2004) interactive alignment theory, which states that primed linguistic material becomes available for interlocutors to use with reduced cognitive effort. In order to straddle the gap between these two research traditions, we conducted a case study in LLC-2 where we explored the social functions that resonance has in discourse (whether it expresses agreement or disagreement) and where priming was operationalized as the time it takes for speakers to respond to the interlocutor's prior turn. The results revealed that (i) resonance was more likely to express disagreement than non-resonance, which we interpreted as being due to the mitigating effect of resonance, and (ii) it also led to faster turn transitions, indicating that priming gives speakers the cognitive tools to counter the temporal pressures of impromptu speech. Therefore, while social motivations encourage speakers to respond early, cognitive mechanisms give them the necessary tools, thus pointing to an intricate interplay between the processes.

Clearly, this case study would not have been possible without the LLC-2 audio material. We discuss two reasons. First, while orthographic transcriptions may reveal whether turn transitions involve gaps or overlaps, they do not provide detailed information about their duration. Yet, there are important differences between, say, slight overlaps, as in (1), and outright interruptions. We used the multimodal annotation tool ELAN (Wittenburg et al., 2006) to gauge these differences and to extract reliable measurements of turn transitions in the data. Second, observation of the results revealed that resonance manifests itself not only at the level of words and structures but also prosodically. More specifically, the original audio file of (1) shows that both utterances carry a rising-falling pitch, suggesting that prosody further contributes to B's desire to mitigate her disagreement with A. Thus, future work could extend the research in this study further by also investigating the role of prosodic resonance in speakers' perceptions of social proximity among themselves. In conclusion, the case study demonstrated the importance of the LLC-2 audio material in extending the scope of spoken corpus linguistics to also include prosodic and temporal investigations of language.

### *References*

Du Bois, J. W. (2014). Towards a dialogic syntax. *Cognitive Linguistics, 25*(3), 359-410.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *TRENDS in Cognitive Sciences, 8*(1), 8-11.

Hirst, D. (2013). Anonymising long sounds for prosodic research. In B. Bigi & D. Hirst (Eds.), *Proceedings of the International Workshop Tools and Resources for the Analysis of Speech Prosody* (pp. 36-37). Laboratoire Parole et Langage.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (Eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556-1559). Genoa: ELRA.

**Rabadán, Rosa**

**University of León**

**rosa.rabadan@unileon.es**

**Ramón, Noelia**

**University of León**

**noelia.ramon@unileon.es**

Type of Contribution: Pre-Conference Workshop Submission

# Semantic and pragmatic annotation of domain-restricted English-Spanish comparable corpora

## A case study

*Abstract*

Richly annotated corpora are essential for the retrieval of usable information in different applied environments. In bilingual corpora, multi-layered annotation becomes essential to carry out detailed contrastive studies, and as a basis for applications in the ever-increasing hybrid, human-machine text production flows. Most bilingual corpora feature at least PoS annotation, and some include other types of lexicogrammatical annotation, e.g. cohesive devices in Kunz & Lapshinova-Koltunski (2018), or genre-specific multiword combinations in Pizarro Sánchez (2017), among others. However, semantically and pragmatically annotated bilingual corpora are still rare and very much in demand. Rarer still are pragmatically annotated corpora of written texts, e. g. Marín-Arrese's CESJD tagset (2017, 2019), or Weisser's proposal for the in-progress TART dataset (Weisser, 2018, p. 280ff).

This paper explores the multi-layer annotation of a domain-restricted English-Spanish comparable corpus focusing on semantic and pragmatic annotation. The starting point is the CLANES corpus compiled in the period 2016-2019. This corpus includes 831,452 words distributed in seven subcorpora corresponding to two different genres: informational-promotional

texts of gourmet foods and drinks and culinary recipes. Initially, the corpus had been rhetorically, and PoS tagged (Labrador et al., 2014). It soon became evident that the use of these materials was limited to contrastive rhetoric and grammatical analyses and that attempts to go beyond these boundaries required higher-level semantic and pragmatic information. Another limitation was the need to conduct unsupervized annotation at these levels on large amounts of texts.

Annotation at the semantic level started from the USAS scheme (UCREL Semantic Analysis System), which was expanded and enriched for Spanish with embedded subcategories in the F1 and F2 sections (Food/Drink). The USAS framework has a multi-tier structure with 21 major discourse fields with the possibility of further fine-grained subdivisions. Our corpus was semantically tagged, both in English and Spanish, taking single words and multi-word expressions (MWEs) separately. These are recurrent, domain productive semantic patterns showing unitary semantics, i.e., they denote a distinctive referential meaning, including culture-bound and metaphoric sequences, e. g. Victoria sandwich, *Pescado a la espalda* (Grilled fish served with a garlic, olive oil and vinegar sauce on top). MWEs were defined using raw lexical items and semantic and PoS tags in each language, and following a human-informed procedure, the recurrent patterns collated to discriminate those who conveyed a distinctive referential meaning. Then, in addition to the 'componential analysis' of their parts, they were assigned a single tag, F1, F1_B1 in our examples above, or F2 in the case of wines and functional teas, e.g. *Rioja*, *Té Kukicha*. The resulting semantic dataset includes over 5,000 domain-specific entries in both languages, plus an additional 10,000 general language entries in Spanish. This semantic information has been crucial to address the subsequent pragmatic tagging of our corpus.

Pragmatic annotation labels are identified through associated semantic and PoS tags in English and Spanish and applied at the sentence level. Our scheme contemplates six categories, namely <state>, <direct>, <suggest>, <recommend>, <praise> and <evidence>.

<State> simply marks the delivery of referential information, and applies to names of products, dishes, etc., enumerations of constituents, e.g.

<STATE>3CARDN1barricasNCZ2_S2mfrepartidasVLadjM1_I2enPREPZ5999 CARDN1botellasNCO2 E5+</STATE> (3 barrels, 999 bottles produced).

<Direct> identifies a directive, an action that is to be carried out to fulfil a goal, as in <DIRECT>StirVBA1.1.1_M1inINX9.2+saltNNA2.1_F1andCCpepper NNF1_L3</DIRECT>.

<Suggest> signals options offered, that may or may not be put into practice, e.g. <SUGGEST>idealADJA5.1+ paraCSUBIZ5 degustarloVLinf X9_X2.4_P1 conPREPZ5sidraNCF2</SUGGEST>(Ideal to taste with cider).

<Recommend> singles out the chosen course of action available among possibilities, as in <RECOMMEND>ItPPN5 'sVBZO2 bestJJSX3.1_X3.5_O4.3_I1 ifINO4.3_O1 youPPI2.2.1useVBPA1.1.1_A1.5 aDTNULL_2 shallowJJA5_A13.7 bakingVBGA2.1_O4.6_F1 dishNNO2_F1<RECOMMEND>

<Praise> consigns the good properties of the product, as perceived intersubjectively, <PRAISE>EsVSfinA3+_L1_Z5_X2.4 unARTZ5_N1_T3_T1.2_Z8bizcochoNCF1 muyADVA13.3esponjosoADJO4.5_O4.1yCCZ5_A1.8+ sinceramenteADVA5.2+_A5.4+_S1.1.3muyADVA13.3 ricoADJX3.1_X3.5_O4.3_I1</PRAISE> (A fluffy, delicious sponge cake).

<Evidence> adds certain factual information about the product, e.g. <EVIDENCE> GanadorNCX9.2+_S7.3 delPDELZ5 premioNCS7.3Cincho NPO2_A6.2+ dePREPZ5 OroNCO12006CODEN1</EVIDENCE>. (Winner of the Cincho de Oro 2006 award)

Both for semantic and pragmatic annotation, manual tagging was effected on a section of the corpus using, first, regular expressions and symbolic analysis. Then word2vec and fastText Machine Learning algorithms were used for unsupervized annotation. Allocating a pragmatic tag on the sole basis of morphological or semantic metadata proved ineffectual, as the rate of success was below 50% for most categories, both in English and in Spanish. Additional restricting strategies were then implemented in an attempt to boost algorithm performance, namely, hybrid contextual patterns, mixing lexical units and metadata, a list of lexical nodes used as restrictors, and a custom stoplist. Querying is done through the Actres Corpus Manager (ACM https://actres.unileon.es/wordpress/?page_id=44&lang=en). This user platform allows the retrieval of information along with the annotation for the three layers of metadata, namely morphological, semantic and pragmatic.

Current results for semantic annotation show an overall degree of success of 84% in Spanish, including MWEs. For English, the hit rate is close to 90%. Pragmatic annotation shows different degrees of success, with an overall success rate of 75% in Spanish and 62.5% in English. If taken by subcorpus, the rate of success for the informational-promotional genre exceeds 70% in Spanish and is near 60% in English. For the instructive genre, the overall results hit 84% in Spanish and just below 65% in English. If taken by pragmatic category, in Spanish, the success rate ranges between 92% for <recommend> and 43.44% for <suggest>. In English, the accuracy ranges between 88 % for <state> and 0% for <evidence>. Results so far evidence that pragmatic tagging suffers from underspecification in both languages, particularly in English. They also suggest that adding up more detailed information on grammatical functions would improve the usefulness of “supporting metadata" for pragmatic annotation.

Work in progress focuses precisely on adding up linguistic information and also on streamlining the contextual rules to improve ML performance. A wealth of studies and applications are possible on these results, the most immediate the design of a pre-editing workbench for bilingual text production in the food-and-drink domain.

## *References*

Kunz, K., & Lapshinova-Koltunski, E. (2018). English vs. German from a textual perspective: Looking inside chain intersection. In S. Ebeling & H. Hasselgård. *Corpora et Comparatio Linguarum: Textual and contextual perspectives. Bergen Language and Linguistics Studies*, *9*(1), 1-22.

Labrador, B., Ramón, N., Alaiz-Moretón, H., & Sanjurjo-González, H. (2014). Rhetorical structure and persuasive language in the subgenre of online advertisements. *English for Specific Purposes, 34*, 38-47.

Marín Arrese, J. (2017). Multifunctionality of evidential expressions in discourse domains and genres. Evidence from cross-linguistic case studies. In J. Marín Arrese, G. Hassler & M. Carretero (Eds.), *Evidentiality revisited: Cognitive grammar, functional and discourse-pragmatic perspectives.* (pp.195-224). Amsterdam: John Benjamins.

Marín-Arrese, J. (2019). CESJD-JMA Tagset for Annotation of Epistemic and Effective Stance Markers [Data set]. Retrieved from: http://corpusnet.unileon.es/assets/uploads/tools/CESJD-TAGSET.pdf.

Pizarro Sánchez, I. (2017). A corpus-based analysis of genre-specific multi-word combinations: Minutes in English and Spanish. In T. Egan & H. Dirdal (Eds), *Cross-linguistic correspondences* (pp. 221-252). Amsterdam: John Benjamins.

Weisser, M. (2018). *How to do corpus pragmatics on pragmatically annotated data: Speech acts and beyond*. Amsterdam: John Benjamins.

**Šebestová, Denisa**

***Charles University***

**denisa.sebestova@ff.cuni.cz**

Type of Contribution: Pre-Conference Workshop Submission

## Prepositional phraseological patterns in Czech and English

Towards a contrastive study resource

*Abstract*

This study aims to design corpus-informed teaching materials for advanced Czech students of English, reflecting differences in native as opposed to non-native phraseologies. It draws on studies suggesting that even advanced L2 learners use a limited repertory of phraseological sequences in ways which differ considerably from native usage (e.g. Granger, 2017; or Hasselgård, 2019, referring back to Hasselgren, 1994, terms these 'phraseological teddy bears'). Limited phraseological choices may hinder the students´ language production in terms of accuracy.

Specifically, I investigate phraseological patterning involving prepositions. As pointed out by Hunston (2008), focus on patterns containing function can efficiently point out textual structure.

The data comes from SYN2015 (Czech) and BNC (English; 100 mw each). A list of the 10 most frequent prepositions was compiled for either corpus. In this work-in-progress report I focus on the corresponding preposition pair *in* - *v*. I extract 3-5-grams containing the preposition in any slot, using *Engrammer* software (Milička 2019). *Engrammer* enables searches for sequences of different lengths at once, comparing their collocation strength. It also reveals alternations with the keyword, detecting pattern variants.

The resulting n-grams are examined to identify prepositional patterns. These are classified by their semantics and textual functions and compared between Czech and English. Results suggest that *in/v* patterns mostly fulfil corresponding functions in the languages compared, although the distribution

of these functions differs. Some pattern classes are only found in English, highlighting its analytic nature as opposed to Czech. Further, while some functional-semantic pattern classes comprise a diverse set of expressions (e.g. adverbials of place), others are limited to few patterns (e.g. emphasizers), suggesting that the corresponding functions are associated with more conventionalized forms of realization.

### *References*

Granger, S. (2017). Academic phraseology: A key ingredient in successful L2 academic literacy. In H. Fjeld & J. Henriksen (Eds.), *Academic language in a Nordic setting: Linguistic and educational perspectives. Oslo Studies in Language, 9*(3), 9-27.

Hasselgård, H. (2019). Phraseological teddy bears: frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In M. Mahlberg & V. Wiegand (Eds.), *Corpus linguistics, context and culture* (pp. 339-362). Berlin: Mouton De Gruyter.

Hunston, S. (2008). Starting with the small words. *International Journal of Corpus Linguistics, 13*(3), 271-295.

Milička, Jiří. (2019). *Engrammer* [software]. Retrieved from: http://milicka.cz/en/engrammer/.