ICAME46

Vilnius University

Book of Abstracts

Per Corpora ad Astra:

Exploring the Past, Mapping the Future

17-21 June 2025

Vilnius

Contents

Keynote speakers

SEBASTIAN HOFFMANN
Universität Trier (Germany)

Sebastian Hoffmann (PhD, University of Zurich) is Professor of English Linguistics at Trier University. Before moving to Trier, he spent three years at Lancaster University (UK) as Lecturer in English Linguistics (2006 - 2009). His research has predominantly focused on the application of usage-based approaches to the study of language and includes topics such as syntactic change/grammaticalization, use and pragmatics of tag questions, and the lexico-grammar of New Englishes. He is a co-author of BNCweb, a user-friendly web-interface to the British National Corpus (BNC), which also forms the basis for his textbook publication Corpus Linguistics with BNCweb – a Practical Guide (Peter Lang, 2008; with S. Evert, N. Smith, D. Lee and Y. Berglund-Prytz). In recent years, his interests have expanded to include corpus phonetics, most prominently on the basis of the Audio BNC.

Audio data in corpus linguistics – challenges and opportunities

Corpus linguistic analysis of audio data has traditionally lived a comparatively shadowy existence at ICAMEs, and this is particularly the case with respect to phonetic and phonological research questions. There is relatively little "conference attendance overlap" with the research community of Corpus Phonetics, for example, and few of the speech corpora (e.g. the Switchboard Corpus or the Buckeye Corpus) used by that community feature in studies by traditional ICAMErs, as these corpora tend to be smaller and more specialised collections of data that are often sampled in controlled contexts to ensure sufficient audio quality. In turn, spoken corpora used by the ICAME community such as ICE-GB or the Santa Barbara Corpus typically do not meet the requirements of corpus phoneticians, for example because they lack phonetic/phonological annotation or sufficient audio quality, or in fact because audio data is not available in the first place (e.g. the London-Lund Corpus or BNC2014).

A major source of audio data for corpus analysis became available when about 5.4 million words of the original BNC were first digitized and phonemically transcribed (see Coleman et al. 2011, Coleman et al. 2011) and later integrated into BNCweb (see Hoffmann & Arndt-Lappe 2021). The data has since been used for a range of corpus(-phonetic) analyses, including my own joint research on intrusive /r/ (Hoffmann & Arndt-Lappe 2021) and stress shift (Arndt-Lappe & Hoffmann 2022). This work has highlighted both the great potential of the data (e.g. size, variety, naturalness) and its drawbacks (e.g. audio quality, issues with forced alignment).

In my talk, I will return to the topic of stress shift, using the AudioBNC for an investigation of the Principle of Rhythmic Alternation ("PRA", Sweet 1876) and the theoretical questions that arise from such an undertaking. In particular, I will present some findings that significantly extend what was presented in Arndt-Lappe & Hoffmann (2022). In doing so, I will provide further evidence that the concept of stress shift must indeed at least partially be questioned, which in turn suggests that some basic tenets of phonological theory may require re-evaluation. I will also discuss some of the methodological challenges that researchers face when using the AudioBNC, but – probably not surprisingly – come to the conclusion that the

advantages by far outweigh these difficulties and that venturing into corpus-phonetic territory is a very worthwhile undertaking for a long-term ICAMEr, too.

References

Arndt-Lappe, Sabine, and Sebastian Hoffmann (2022), "Comparing approaches to phonological and orthographic corpus formats: Revisiting the Principle of Rhythmic Alternation", in Ole Schützler & Julia Schlüter (eds), *Comparative Approaches to Data and Methods in Corpus Linguistics*, Cambridge: Cambridge University Press, 46–72.

Coleman, John, Ladan Baghai-Ravary, John Pybus, and Sergio Grau (2012), "Audio BNC: The Audio Edition of the Spoken British National Corpus", Oxford: Phonetics Laboratory, University of Oxford.

Coleman, John, Mark Y. Liberman, Greg Kochanski, Lou Burnard and Jiahong Yuan (2011), "Mining a Year of Speech. Paper Presented at VLSP 2011: New Tools and Methods for Very-Large-Scale Phonetics Research", University of Pennsylvania, 29–31 January 2011.

Hoffmann, Sebastian, and Sabine Arndt-Lappe (2021), "Better Data for More Researchers: Using the Audio Features of BNCweb", *ICAME Journal*, 45, 125–54.

Sweet, Henry (1876), "Words, Logic, and Grammar", in *Transactions of the Philological Society*, *1875–1876*, 470–503.

ROSA LORÉS
Universidad de Zaragoza (Spain)

Rosa Lorés is Professor of Applied Linguistics at the Universidad de Zaragoza (Spain). She has conducted research on specialized discourse and written academic genres from the standpoint of pragmatics, contrastive rhetoric, genre analysis, and corpus studies. Her current research interests focus on digital scientific communication and dissemination practices. She has co-edited several books and special issues, and her articles have been published in international journals including *English for Specific Purposes, Journal of Pragmatics, Discourse, Context & Media, and Text&Talk).* She co-coordinates the research group InterGEDI (Interpersonality in Digital Genres). She is also a member of the research group CIRES (Intercultural Communication and Societal Challenges), and the Research Institute of Employment, Digital Society, and Sustainability (IEDIS).

Bridging the gap: From experts to audiences in digital science communication

Institutions and scientific organizations have actively promoted the popularization of science since the mid-20th century. As a result, the notion of a strict divide between the scientific community and the lay public has been challenged (Jones et al., 2015; Pilkington, 2018; Freddi, 2020), leading to the understanding that popular science is crafted for diverse audiences, including experts, with varying degrees of background knowledge, needs, and expectations (Myers, 2003; Hyland, 2010). Thus, the reader "is no longer a passive observer but an active participant in the social discourse related to science and its consequences" (Pilkington, 2018: 14).

In this talk I will examine the digital communication of scientific knowledge, focusing on how digital technology has transformed the dissemination of information to provide alternative research methods, collaboration, and discursive practices. Central phenomena associated with digital knowledge such as multimodality, interactivity, and recontextualization, as well as their implications for bridging knowledge asymmetries will also be discussed.

A theoretical and methodological framework for digital discourse analysis is proposed, integrating genre studies, corpus studies, and pragmatics. The combination of these methodologies with the SciDis Database (Pascual and Sancho-Ortiz, 2025) allows researchers to conduct in-depth analyses of scientific dissemination practices across different disciplines, modes and media. The SciDis database is presented as a tool to explore digital scientific practices related to areas of current interest such as health, economy, and natural sciences. The database identifies and classifies various digital discursive practices, considering variables such as typology (web-hosted vs. social media practices) and content generator (author-generated vs. writer-mediated practices). A study derived from this database will be presented by way of

illustration. This combination of theoretical and methodological perspectives provides a solid foundation for future studies on the evolving landscape of digital scientific communication and its impact on knowledge transfer in the digital age.

References

Freddi, Maria (2020), Blurring the lines between genres and audiences: Interaction in science blogs, *Discourse and Interaction* 13(2), 9–35. https://doi.org/10.5817/DI2020-2-9

Hyland, Ken (2010), Constructing proximity: Relating to readers in popular and professional science, *Journal of English for Academic Purposes* 9(2), 116–127. https://doi.org/10.1016/j.jeap.2010.02.00

Jones, Rodney H., Chik, Alice and Christoph A. Hafner (2015), Introduction. Discourse analysis and digital practices, in Rodney H. Jones, Alice Chik and Christoph A. Hafner (eds), *Discourse and Digital Practices. Doing Discourse Analysis in the Digital Age*, Routledge, 1–17.

Myers, Greg (2003), Discourse studies of scientific popularization: questioning the boundaries. *Discourse Studies* 5(2), 265–279. https://doi.org/10.1177/1461445603005002006

Pascual, Daniel, and Ana Eugenia Sancho-Ortiz (2025), Investigating recontextualisation processes in scientific digital practices: The SciDis Database, *Revista Electrónica de Lingüística Aplicada* 23, 101–118. https://doi.org/10.58859/rael.v23i1.649

Pilkington, Olga (2018), *Presented Discourse in Popular Science: Professional Voices in Books for Lay Audiences*, BRILL.

RŪTA PETRAUSKAITĖ
Vytautas Magnus University (Lithuania)

Rūta Petrauskaitė is a professor at the Department of Lithuanian Studies. Currently she acts as the director of the Institute of the Digital Resources and Interdisciplinary Research (SITTI) at Vytautas Magnus University.
In the last decade she has been a vice-president of the Research Council of Lithuania and the Chair of the Committee of Social Sciences and Humanities. Internationally she got involved in the activities of the Common Language Resources and Technology Infrastructure (CLARIN), Science Europe Research Data working group and European Open Science Cloud (EOSC).
Her research interests comprise a range of topics from linguistics to discourse analyses. She initiated and supervised compilations of the first big corpora of the Lithuanian language and corpus-based research in a few fields of linguistics.
Rūta Petrauskaitė is a proponent of data-driven research, Open Science and data sharing initiatives.
https://www.vdu.lt/cris/entities/person/ruta-petrauskaite

Corpora and data. From John Sinclair to artificial intelligence

Last year we celebrated thirty years anniversary of corpus linguistics in Lithuania. The advent of the new trend was gradual, nevertheless, groundbreaking. Our participation in EU projects TELRI I and TELRI II speeded up compilation of the first corpora for the Lithuanian language that was followed by corpus-based research. To deal with corpora we badly needed new methodological approaches, happily, by that time they were already available in publications by John Sinclair as well as his activities related to COBUILD. TELRI was beneficial due to co-operation with linguists from other countries but most of all due to the revolutionary ideas and personality of John Sinclair.

John Sinclair was ahead of time in his attempts to describe how meaning is created in human language. His holistic approach is based on a few key concepts of lexical items juxtaposed to ortographic words or extended units of meaning, comprising elements of lexis (collocation), grammar (colligation), semantics (semantic preference) and pragmatics (attitudinal meaning). His effort to do away with the historical split of lexis and grammar and to show the close relation between the two types of patterns more than thirty year ago was truly astonishing.

Main cornerstones of his language theory included: a) reunification of grammar as structure and lexis as vocabularly for a language for creation of meaning in text, i.e., form and meaning in language that cannot be separated; b) the importance of co-text and context for generating and understanding the meaning; c) reliance on corpora as large amounts of language data for pattern detection instead of testing hypothesis, i.e., corpus-driven instead of corpus based approach; d) reluctance to trust man-made consensus grammar based annotation.

John Sinclair passed away in 2007, before neuronic revolution so he did not witness its main developments that went along the same lines as he suggested for corpus linguistics. Major steps in the

direction of AI were as follows: 1990 marked the shift from rule- to statistics-based methods and machine learning. 2014 brought neuronic language technologies, that caused a major paradigm shift in natural language processing, specifically the shift from rule-based approaches to data-driven approaches. The focus has increasingly moved toward high-quality corpus modelling rather than relying on explicit grammar rules or predefined linguistic annotations. Large language models like GPT, released three years ago represent this evolution: they were fundamentally data-driven but increasingly incorporating techniques to inject linguistic knowledge and structure where it is beneficial. High-quality corpora enabled models to learn language patterns effectively and this is how AI learned languages – by encompassing broad co-text and capturing the richness and complexity of natural language.

LUKAS SÖNNING
Universität Bamberg (Germany)

Lukas Sönning is a post-doctoral researcher associated with the Chair of English Linguistics at the University of Bamberg (Germany). Following his PhD project, which looked at phonological features in German Learner English, his interest shifted to statistical aspects of corpus-linguistic methodology. He has worked on topics such as keyness analysis, dispersion, and down-sampling, and his habilitation (post-doc) project concentrates on the linguistically grounded use of mixed-effects models in variationist corpus research.

Lukas has also been an active promoter of open-science practices and his work is strongly informed by his passion for data visualization. He is currently also involved in a DFG-funded project on the analysis of high-dimensional survey data drawn from the BSLVC (Bamberg Survey of Language Variation and Change).

Per corpora et diagrammata ad astra: Data visualization in corpus linguistics

Since corpus-based work often involves the quantitative analysis of relatively complex data sets, data visualization has always played a critical role in our field. Today we are confronted with an unprecedented supply of graph types, which are in many cases relatively straightforward to implement with freely available software such as R. While this overabundance holds out many opportunities both for the individual researcher and for the scientific community, it also necessitates critical reflection and debate about the merits and added value of (novel) graph types.

This talk traces the evolution of corpus data visualization over the course of the past 30 years. An analysis of just over 1,200 published corpus-linguistic research articles allows us to chart emerging practices in the field, identify trends, and examine the state-of-the-art. We observe that the usage rate of graphs has increased over time, and, as a means of data communication, they are nowadays on a par with tabular displays. While our review does detect a recent influx of novel graph types, the usage rate of traditional forms is remarkably stable over time, suggesting that certain workhorses of data visualization are here to stay.

The present talk will illustrate what a constructive discourse about data visualization in our field could look like. To this end, I will examine the use of the three most common graph types – bar charts (37% of articles), line plots (23%), and scatterplots (14%) – from the viewpoint of the design recommendations given in the data visualization literature. This kind of critical review allows us to see where we stand, and to acknowledge room for improvement. As this discussion targets the common core of visuals in corpus linguistics, it is likely to be of relevance for most practicing corpus linguists. Further, we will take a closer look at a number of newcomers, which have recently entered the scene of presentation graphs in corpus-linguistic journals. Specifically, we will examine (the use of) dendrograms, mosaic charts, and CARTs (classification and regression trees) from the perspective of graph construction and perception. Since each

of these forms has applied emphatically for a permanent position in the corpus-linguistic visualization toolbox, a careful engagement with their (potential) weaknesses is needed.

This talk will give us an opportunity to take a dive into the fascinating field of visualization research, including its cognitive underpinnings and empirical grounding. It will be apparent that there are many implicit and explicit parallels between work on (statistical) data visualization and the study of language.

Pre-conference Workshops

Workshop 1: Speech acts in formal and informal interactions in English: Mapping the past, exploring the future

Convenors:
Rachele De Felice (The Open University)
Christine Elsweiler (University of Innsbruck)
Sofia Rüdiger (FU Berlin)

In recent decades, speech act use in English has received considerable attention, producing a substantial body of research drawing on both historical and present-day data. Studies have covered a range of speech acts, including, among others, apologies, compliments, requests and thanking. These have been explored across different types of interactions and communicative settings, such as authentic and fictional conversations (e.g., Schauer and Adolphs 2006, Culpeper and Archer 2008, Jucker and Taavitsainen 2008, Jucker et al. 2008, Jautz 2013, Jucker 2017, Haselow 2024), letter-writing, email correspondence and other types of computer-mediated communication (e.g., Lutzky and Kehoe 2017, Murphy and De Felice 2018, De Felice and Moreton 2019, Elsweiler 2024), courtroom discourse (Archer 2005, Kryk-Kastovsky 2009, Chaemsaithong 2018) as well as service encounters and other workplace interactions (e.g., Bös 2007, Vine 2009, Fox and Heinemann 2021, Barron 2022), to name but a select few.

This body of work, drawing on spoken and written as well as naturally occurring and elicited data, has enabled us to develop an empirically-grounded understanding of speech act use across various forms of interaction, moving beyond their initial descriptions founded on individual intuitions. At the same time, the fact that this work is based on a wide range of sources and often (understandably) reflects specific research interests of individual scholars limits the comparability of findings for particular speech acts. This is the result of a number of factors, including – but not limited to – the use of different categorisation approaches to speech acts, form-first vs. function-first searches, availability of speaker metadata, and types of texts included in the corpora.

In this workshop, we therefore aim to tackle the issue of comparability of corpus-based speech act research. Our discussions will be informed by three key research questions:

1. How can we make different datasets comparable?
2. How can we draw overarching conclusions from individual studies?
3. Is it possible to draw diachronic comparisons if our knowledge about communicative settings is limited?

We invite contributions on speech act use in either formal or informal interactions drawing on historical or present-day English corpus data, or both, that address the issue of comparability by focusing on methodological aspects relating to

- which data are used
- the size of datasets
- how and whether datasets are prepared or annotated
- which metadata are available
- how conventional and non-conventional speech acts are retrieved in non-annotated data.

We are additionally interested in contributions which explicitly compare corpus-based speech act research to insights from other methodological approaches such as experimental studies, elicitation studies, etc.

References

Archer, Dawn (2005), *Questions and Answers in the Historical Courtroom (1640-1760): A Sociopragmatic Analysis*, Amsterdam and Philadelphia: John Benjamins.

Barron, Anne (2022), "'Sorry Miss, I completely forgot about it': Apologies and Vocatives in Ireland and England", in Stephen Lucek and Carolina P. Amador-Moreno (eds), *Expanding the Landscapes of Irish English Research: Papers in Honour of Dr. Jeffrey J. Kallen*, New York: Routledge, 109–128.

Bös, Birte (2007), "What do you lacke? What is it you buy? Early Modern English Service Encounters", in Susan Fitzmaurice and Irma Taavitsainen (eds), *Methods in Historical Pragmatics*, Berlin: De Gruyter Mouton, 219–240.

Chaemsaithong, Krisda (2018), "Investigating Audience Orientation in Courtroom Communication: The Case of the Closing Argument", *Pragmatics and Society* 9(4), 545–570.

De Felice, Rachele and Emma Moreton (2019), "Identifying Speech Acts in a Corpus of Historical Migrant Correspondence", *Studia Neophilologica* 91(2), 154–174.

Elsweiler, Christine (2024), "Modal May in Requests: A Comparison of Regional Pragmatic Variation in Early Modern Scottish and English Correspondence", *Journal of Historical Pragmatics* 25(3), 355–391.

Fox, Barbara A. and Trine Heinemann (2021), "Are They Requests? An Exploration of Declaratives of Trouble in Service Encounters", *Research on Language and Social Interaction* 54(1), 20–38.

Haselow, Alexander (2024), "Politeness, Speech acts and Socio-cultural Change: The Expression of Gratitude in the History of English", *Journal of Historical Pragmatics* 25(3), 419–449.

Jautz, Sabine (2013), *Thanking Formulae in English: Explorations across Varieties and Genres*, Amsterdam and Philadelphia: Benjamins.

Jucker, Andreas H. and Irma Taavitsainen (2008), "Apologies in the History of English: Routinized and Lexicalized Expressions of Responsibility and Regret", in Andreas H. Jucker and Irma Taavitsainen (eds), *Speech Acts in the History of English*, Amsterdam and Philadelphia: Benjamins, 229–244.

Jucker, Andreas H., Gerold Schneider, Irma Taavitsainen and Barb Breustedt (2008), "Fishing for Compliments: Precision and Recall in Corpus-linguistic Compliment Research", *Speech Acts in the History of English*, Amsterdam and Philadelphia: Benjamins, 273–294.

Kryk-Kastovsky, Barbara (2009), "Speech Acts in Early Modern English Court Trials", *Journal of Pragmatics* 41(3), 440–457.

Lutzky, Ursula and Andrew Kehoe (2017), "'I Apologise for my Poor Blogging': Searching for Apologies in the Birmingham Blog Corpus", *Corpus Pragmatics* 1(1), 37–56.

Murphy, M. Lynne and Rachele de Felice (2018), "Routine Politeness in American and British English Requests: Use and Non-use of Please", *Journal of Politeness Research* 15(1), 77–100.

Schauer, Gila and Svenja Adolphs (2006), "Expressions of Gratitude in Corpus and DCT Data: Vocabulary, Formulaic sequences, and Pedagogy", *System* 34(1), 119–134.

Taavitsainen, Irma and Andreas H. Jucker (2008), "'Methinks You Seem More Beautiful than Ever': Compliments and Gender in the History of English", in Andreas H. Jucker and Irma Taavitsainen (eds), *Speech Acts in the History of English*, Amsterdam and Philadelphia: Benjamins, 203–228.

Vine, Bernadette (2009), "Directives at Work: Exploring the Contextual Complexity of Workplace Directives", *Journal of Pragmatics* 41(7), 1395–1405.

A corpus-pragmatic function-based investigation of questions and answers in contemporary spoken British and American English

Dawn Archer and Niall Curry
(Manchester Metropolitan University)

Questions and answers play a central role in spoken communication. As a rhetorical "talk-in interaction" device, questions can help speakers to manage turn-taking, demonstrate engagement, and give the communicative "floor" to interlocutors (Curry & Mark 2025). Questions and answers have been studied in casual as well as strategic contexts (the courtroom, classroom, etc.), in historical and modern periods, across real-world activity types, in fictional contexts, and across languages (Archer 2005, 2012; Atkinson & Drew 1979; Bolden et al. 2023; Peñarroja, 2020; Sinclair and Coulthard 1975). Questions in casual spoken communication often occur at the close of turns and, in the case of polar questions, are typically followed by mitigating devices ('well', 'I mean', 'I don't know') rather than a yes or no response (Curry & Mark 2025) as a means of softening the discourse. Taxonomically, questions are identified via their subject-verb

inversion or rising intonation and are typically identifiable in transcripts through the Illocutionary force indicating device (IFID) of the question mark (Flöck & Geluykens 2015). Answers are then (normally) identified as second-pair parts to questions (but see Sinclair and Coulthard 1975). Studies of questions and answers tend to adopt a manual or semi-automatic approach to identification. Limited attention has been paid, in the extant literature, to indirect questions and answers in spoken language, however, despite their evident value, in English, as a less face-threatening approach to conversation management and interlocutor engagement. This paper is part of a small body of work interested in (in)validating an automatic approach to the effective retrieval of such phenomena (see, e.g., Landert et al 2023; Jucker 2024). We investigate questions and answers in spoken English conversation, using comparable American and British corpora. With regard to the paper's methodological contribution, we first assess the extent to which (in)direct questions can be found automatically using Curry's (2021; 2023) IFID approach on the two corpora. We then draw upon Archer's (2005) taxonomy of answer types – a taxonomy that was applied manually for semi-automatic interrogation – to determine whether it is possible to effectively categorize responses to these IFIDs. This allows us to confirm the roles of questions and answers; including the extent to which some answers can carry "(one or more) of several illocutionary force(s) at any given time, and still function as an answer", as Archer (ibid: 290) claims. The work is thus designed to further develop our theoretical understanding of answers (in relation to questions) as well as enabling us to assess the benefits and limitations of automatic pragmatic annotation (Archer et al. 2008; Lu 2014). An additional theoretical advantage is that the inter-varietal analysis, contrasting both American and British varieties, and the intra-varietal analysis, documenting regional variation within American and British contexts, offers insight into evident lacunae in the wider literature on casual spoken language, i.e., the shared and differing questioning and answering practices in casual spoken conversation in American and British Englishes.

References

Archer, Dawn (2005), *Questions and Answers in the Historical Courtroom (1640-1760)*, Amsterdam and Philadelphia: John Benjamins.

Archer, Dawn (2012), 'Assessing Garrow's *aggressive* questioning style', in Gabriella Mazzon (ed), *English Historical Dialogue Studies,* Milano: FrancoAngeli, 301–320.

Archer, Dawn, Jonathan Culpeper and Matthew Davies (2008), Pragmatic annotation, in Anke Lüdeling and Merja Kytö (eds), *Corpus Linguistics: An International Handbook*, Mouton de Gruyter, 613–641.

Atkinson, J. Maxwell and Paul Drew (1979), *Order in Court: The Organisation of Verbal Interaction in Judicial Settings,* London: Macmillan.

Bolden, Galina B., John Heritage and Marja-Leena Sorjonen (2023), *Responding to Polar Questions Across Languages and Contexts*, Amsterdam and Philadelphia: John Benjamins.

Curry, Niall (2021), *Academic writing and reader engagement: Contrasting questions in English, French and Spanish corpora,* London: Routledge. https://doi.org/10.4324/9780429322921

Curry, Niall (2023), Question illocutionary force indicating devices in academic writing: A corpus-pragmatic and contrastive approach to identifying and analysing direct and indirect questions in English, French, and Spanish, *International Journal of Corpus Linguistics*, 28(1), 91–119. https://doi.org/10.1075/ijcl.20065.cur

Curry, Niall and Geraldine Mark (accepted). Applications of corpus linguistics in language education: teacher, editor, and assessment developer perspectives, *Research Notes*.

Flöck, Ilka and Ronald Geluykens (2015), Speech acts in corpus pragmatics: A quantitative contrastive study of directives in spontaneous and elicited discourse, in J. Romero-Trillo (ed), *Yearbook of corpus linguistics and pragmatics,* London: Springer, 7–37. https://doi.org/10.1007/978-3-319-17948-3_2

Jucker, Andreas H. (2024), *Speech Acts: Discursive, Multimodal, Diachronic*, Cambridge: Cambridge University Press.

Landert, Daniela, Daria Dayter, Thomas C. Messerli and Miriam A. Locher (2023), *Corpus Pragmatics,* Cambridge: Cambridge University Press.

Lu, Xiaofei (2014), *Computational Methods for Corpus Annotation and Analysis*, Netherlands: Springer.

Peñarroja, Manuel Rodríguez (2020), *Analysing the Pragmatics of Speech Acts in Sitcom and Drama Audiovisual Genres,* Newcastle: Cambridge Scholars.

Sinclair, John McHardy and Malcolm Coulthard (1975), *Towards and Analysis of Discourse: The English Used by Teachers and Pupils,* Oxford: Oxford University Press.

Automatic prediction of speech acts

Sophia Conrad, Tilia Ellendorff and Gerold Schneider
(University of Zurich)

The corpus of spoken Irish English, SPICE Ireland [1] has been manually annotated for the five speech act categories of illocutionary speech acts, following [4], namely *representatives*, *directives*, *commissives*, *expressives*, and *declaratives*. Using this annotated dataset as training data, we use state-of-the-art machine learning approaches to predict speech acts.

We address the following four research questions (RQ):

1. How well does a classical document classifier perform on the task of speechact classification?
2. To what extent does class imbalance affect classification performance, and what strategies can mitigate the overrepresentation of large classes (e.g. *representatives*) compared to smaller classes (e.g. *declaratives*)?
3. Do distributional semantics approaches, such as word and sentence embeddings, improve speech act classification compared to traditional wordbased methods?
4. Can large language models (LLMs) outperform traditional machine learning approaches in speech act classification?

Our level of annotation are the turns, as given in SPICE Ireland. The key features used in our approach include the vectorized text of the current turn, as well as the preceding and following turns for context. Additionally, we incorporate register information at three levels as given by the corpus. For some experiments, we also include the gold label of the previous speech act.

Concerning RQ1, we report preliminary results on two baselines, namely logistic regression (given in Table 1, "Baseline F-Score") and a rule-based system.

Performance on small classes is considerably worse than on large classes (see RQ2). In order to boost underrepresented classes, we thus test oversampling strategies. To this end, we employ GPT-4 [2] as a data augmentation tool to create further instances of small classes (marked with * in Table 1), providing the model with some examples from the corpus and prompting it to generate

Class	Class Description	Frequency	Baseline F-Score	Augmented F-Score
rep	Representatives	35246	0.84	0.84
dir	Directives	10522	0.57	0.44
icu	Indet.communicativeunits	3782	0.76	0.27
xpa	Notanalysable	2513	0	0.26
exp*	Expressives	1242	0.01	0.17
com*	Commissives	685	0.04	0.56
soc*	Socialgreetings	173	0.39	0.08
dec*	Declaratives	57	0	0.02
D	Micro-AverageAccuracy	54220	0.72	0.73

Table 1: Comparison of Baseline System to Augmented System

similar turns. The performance of a system using a more balanced sample and parameter tuning is given in Table 1, "Augmented F-Score"). There is only improvement in some classes while partially impairing the performance of the larger classes.

Speech acts are pragmatic categories and thus often expressed very indirectly. Accordingly, turning to RQ3, we can expect on the one hand that approaches including distributional semantics (ranging from word embeddings via document embeddings, e.g. Roberta or snowflake, to GPT-4 and other large language models) leverage contextual information. On the other hand, pragmatics as the art of reading between the lines and understanding situational knowledge may still remain a challenging task for automatic systems which cannot understand texts in a human sense [3]. Preliminary results comparing a parameter-tuned (but not data augmented) system, show that we could increase accuracy from 0.76 to 0.79 with *word2vec* word embedding, indicating that embeddings partly manage to include situational knowledge.

References

Kallen, Jeffrey L. and John Kirk (2012), *SPICE-Ireland: A user's guide*. Cló Ollscoil na Banríona.

OpenAI et al. (2024), *GPT-4 Technical Report*. arXiv: 2303.08774[cs.CL]. https://arxiv.org/abs/2303.08774.

Schneider, Gerold (2024), Automatically detecting Directives with SPICE-Ireland, in Martin Schweinberger and Patricia Ronan (eds), *Sociopragmatic Variation in Ireland: Using Pragmatic Variation to Construct Social Identity*. Berlin, Boston: De Gruyter Mouton.

Searle, John R. and Jerrold M. Sadock (1976), Toward a linguistic theory of speech acts, *Language*, 52(4), 966.

Apology conventions in business-related correspondence from the 18th century to the present day

Christine Elsweiler and Rachele De Felice
(University of Innsbruck, The Open University)

PDE apologies are highly routinised speech acts which are typically realised explicitly by the IFID *sorry* (Deutschmann 2003). However, it is also possible to apologise indirectly through other strategies such as taking on responsibility for an offense or by combining apology strategies (Holmes 1990: 167; Aijmer 1996: 82-84). In contrast to Present-Day English, the speaker-oriented IFID *sorry* was less prominent than addressee-oriented forms such as *(I pray you) pardon me* in Late Modern English and combinations of indirect apology strategies are often found. Apologies were, moreover, less routinised and more complex (Jacobsson 2004; Jucker & Taavitsainen 2008; Jucker 2018). Nevertheless, in genres such as letter-writing, certain conventional apology patterns are found, for instance, the recurrent *trouble* formula, e.g., *yet I cannot but giue you this trouble*. In addition to these general trends, the choice of apology strategies also depends on the type of offense that is being apologised for, e.g., time offenses such as being late or talk offenses such as interrupting someone (Aijmer 1996: 109; Deutschmann 2003: 62, Thaler & Elsweiler 2023: 234) as well as the communicative setting (Thaler & Elsweiler 2023).

In this paper, we address apologies in the communicative setting of British English business-related correspondence from a diachronic perspective. Our goals are

1. to examine if offense types in business-related correspondence, for instance, slowness in replying or length of the letter/e-mail, are stable from the Late Modern period through to the present day,
2. to explore which apology strategies and combinations of strategies are chosen for the different offense types and how these are formally realised across the periods under investigation.

Our data are drawn from different correspondence corpora and editions spanning the 18th to the 21st century, totaling between 100 and 150 letters per century. For the 18th century, they comprise Scottish letters on estate business from the *Helsinki Corpus of Scottish Correspondence, 1540–1750* (ScotsCorr), the correspondence of the Scottish philosophers Adam Fergusson and Adam Smith with their book-sellers, printers and agents, as well as business-related letters from the *Corpus of Early English Correspondence Extension Sampler* part 1 and 2 (CEECES1 & CEECES2). The 19th- and 20th-century data include letters from the *British Telecom Correspondence Corpus* (BTCC; Morton and Nesi 2019) and the professional letters sub-component of the 1994 *British National Corpus* (BNC1994; Burnard 2000). For the 21st century, we look at the publicly available emails recently released as part of the Post Office Horizon

Inquiry (https://www.postofficehorizoninquiry.org.uk/). The apologies are retrieved combining a form-to-function approach with close reading. Each instance is categorised for both formal features (i.e. strategies used) and offence types (cf. Ancarno 2005, De Felice 2024, Elsweiler 2024). Previous research on other datasets has found that minor offences centred around correspondence- and communication-related issues are the most frequent categories in emails and that they do not typically require complex apology strategies (Harrison and Allton 2013, Marsden 2019, De Felice 2024). Our work will compare whether these findings persist across different datasets and time periods and can therefore be considered a constant feature of correspondence.

References

Aijmer, Karin (1996), *Conversational Routines in English: Convention and Creativity*, London/New York: Longman.

Ancarno, Clyde (2005), The Style of Academic E-mails and Conventional Letters: Contrastive Analysis of Four Conversational Routines, *IBÉRICA* 9, 103-122.

Burnard, Lou (2000), *Reference Guide for the British National Corpus*, Oxford: Oxford University Press.

CEECES 1 = *Corpus of Early English Correspondence Extension Sampler* part 1 (2021), Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja S.ily and Anni Sairio at the Department of Languages, University of Helsinki, XML conversion by Lassi Saario, Helsinki: VARIENG.

CEECES 2 = *Corpus of Early English Correspondence Extension Sampler* part 2 (2022), Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio at the Department of Languages, University of Helsinki, XML conversion and encoding by Lassi Saario, Helsinki: VARIENG.

De Felice, Rachele (2024), "Stability of Pragmatic Markers: The Case of *Sorry* in Organizational Emails from the Clinton Email Corpus", paper presented at the *Sociopragmatic Variation in Late Modern English Workshop*, ICAME45, Vigo, Spain.

Deutschmann, Mats (2003), *Apologising in British English,* Umeå: Skrifter Från Moderna Språk.

Dieter Stein, Tuija Virtanen (eds), *Pragmatics of Computer-Mediated Communication*, Berlin: Walter de Gruyter, 315–337.

Elsweiler, Christine (2024), "The Organisation of Macro-requests in Early 18th-century Scottish and English Letters", paper presented at the *Sociopragmatic Variation in Late Modern English Workshop*, ICAME45, Vigo, Spain.

Harrison, Sandra, and Diane Allton (2013), Apologies in Email Discussions, in S. Herring, D. Stein and T. Virtanen (eds), *Pragmatics of Computer-Mediated Communication*, Berlin/Boston: De Gruyter, 315–338.

Holmes, Janet (1990), Apologies in New Zealand English, *Language in Society,* 19, 155–99.

Jacobsson, Mattias (2004), Apologies and Apologetic Attitude in Early Modern English, *Nordic Journal of English Studies,* 3(3), 187–204.

Jucker, Andreas H. (2018), Apologies in the History of English: Evidence from the *Corpus of Historical American English* (COHA), *Corpus Pragmatics,* 2(4), 375–398.

Jucker, Andreas H. and Irma Taavitsainen (2008), "Apologies in the History of English: Routinized and Lexicalized Expressions of Responsibility and Regret," in Andreas H. Jucker and Irma Taavitsainen (eds), *Speech Acts in the History of English*, Amsterdam/Philadelphia: Benjamins, 229–244.

Marsden, Elizabeth (2019), *Relationship Management in Intercultural Business Emails*, Doctoral thesis, University of Huddersfield, available at http://eprints.hud.ac.uk/id/eprint/34902/.

Morton, Ralph and Hilary Nesi (2019), Institutional Collaboration in the Creation of Digital Linguistic Resources: The Case of the *British Telecom Correspondence Corpus*, in Simon Popple, Andrew Prescott A. and Daniel H. Mutibwa (eds), *Communities, Archives and New Collaborative Practices*, Bristol: Policy Press.

ScotsCorr = *The Helsinki Corpus of Scottish Correspondence 1540-1750* (2017), Anneli Meurman-Solin (ed), Helsinki: University of Helsinki. http://urn.fi/urn:nbn:fi:lb-201411071.

Thaler, Marion and Christine Elsweiler (2023), The Role of Gender in the Realisation of Apologies in Local Council Meetings: A Variational Pragmatic Approach in British and New Zealand English, in Anna Islentyeva and Anatol Stefanowitsch (eds), Special Issue on Gender and Language, *Zeitschrift für Anglistik und Amerikanistik,* 71(3), 217–239.

Speech acts, communicative setting and discursive context: Problems of data and methodology in historical speech act analysis

Alexander Haselow, Julian Häde and Johannes Lässig
(University of Wuppertal)

This talk discusses findings deriving from the *Diachrony of Communicative Actions Project* (DiCAP), which is devoted to the empirical, largely qualitative study of the long-term diachrony of – first and foremost – expressive, and commissive speech acts in the history of English, relating historical-linguistic findings to major transformations in the socio-cultural history of Britain. In this project, verbal actions are identified manually through close inspection of historical corpora of English and collected in a database, paying close attention to their sequential context, if available. As a second step, these actions are coded for contextual variables by which they are conditioned, such as the relationship between the intra- or extradiegetic co-participants, or the setting (e.g. *religious*, *military*, *private*), and according to functional and formal aspects.

In this talk, we will outline the discursive approach underlying the project (see also e.g. Jucker 2016), which studies speech acts not with reference to isolated utterances taken out of their concrete discourse contexts. Rather, it links linguistic observation to information about the way in which specific forms of discourse unfold, to the setting, and to the relation between the interlocutors. We will illustrate how this approach may tackle the following notorious problems in historical speech act analysis:

i. Speech acts are licensed by the discursive context and the setting. Hence, the communicative function of an utterance is not transdiscursively generalisable and, in theory, each utterance requires case-by-case treatment.

ii. Many utterances in historical sources are not embedded in dialogues, yet we have to assume that they carried some pragmatic function, which has to be identified somehow.

iii. Speech acts may occur in sequence and often serve as vehicles for other speech acts.

iv. Speech acts are often hybrid entities, combining different illocutions.

We utilize the discursive approach and, where applicable, the next-action proof-procedure as a supplementary analytical tool for objectively identifying potential illocutions. Using EXPRESSIONS OF GRATITUDE, OF FUTURE COMMITMENT and INVOCATION OF HARM, we will show how access to the communicative context helps to understand important, yet underexplored aspects of the long-term developments of individual verbal actions, focusing on:

a. how the setting by which an utterance may be licensed has changed (e.g. Which factors condition the expression of gratitude, or make it normatively required?, Haselow 2024),

b. how cultural perceptions of certain utterances have changed (manifesting themselves, e.g. in the responses to or in metapragmatic comments on an action, e.g. Brinton 2021),

c.how utterances that once indicated the illocutionary point of one action have to come to serve other actions and thus shifted the domain (e.g. when utterances originally used to perform commissive acts shifted into the domain of expressive acts serving politeness; or when thanking shifted from an act of deference to an act that enhances one's own face in the 16th century),

d. how the utterance forms associated with one speech act have changed over time.

References

Brinton, Laurel (2021), Responding to thanks. From *you're welcome* to *you bet. Journal of Historical Pragmatics,* 22(2), 180–201.

Haselow, Alexander (2024), Politeness, speech acts and socio-cultural change. The expression of gratitude in the history of English, *Journal of Historical Pragmatics,* 25(3), 419–449.

Jucker, Andreas H. (2016), Politeness in eighteenth-century drama: A discursive approach, *Journal of Politeness Research,* 12(1), 95–115.

"Pray pardon the hast this is ended in"
A corpus-based analysis of apologies in Early Modern English correspondence

Daniela Landert and Andreas H. Jucker
(Heidelberg University)

Apologies are usually defined as expressions of regret for a past event for which the apologizer accepts at least some responsibility, but recent work on the history of apologies in English has shown that such definitions need to be treated with care. In the course of time not only the manifestations of apologies have been subject to change but at least to some extent also their functional profile (see Jucker and Taavitsainen 2008; Williams 2018; Jucker 2019). Jucker (2019) proposed the term "attenuation" to account for the weakening of the illocutionary force of apologies over time. Haselow (2024), working on expressions of gratitude, put this into a larger context and suggested the four related processes of recontextualization, functional expansion, attenuation and routinisation to account for the typical development of speech acts across time.

In this contribution, we want to put Haselow's concepts to the test by a focused analysis of apologies in a corpus of relatively informal interactions in Early English. For this purpose, we use the *Parsed Corpus of Early English Correspondence*, consisting of about 2.1 million words spread over five time periods from 1350 to 1710. In a first step, we extract a carefully stratified sample from the entire corpus and manually annotate all attested examples of apologies. This is expected to retrieve all manifestations of apologies that appear with a reasonable frequency in the entire corpus. In a second step, these manifestations are then searched for with appropriate corpus tools to ascertain their distribution in the entire corpus.

Preliminary results indicate that there is an overall increase of apologies from the late fourteenth to the early eighteenth century with clear evidence for Haselow's concepts of functional expansion, attenuation and routinisation. While early instances of apologies consist largely of implicit expressions of regret by the apologizer for a past event and their own responsibility for it, later instances show an increasing level of attenuation and routinized formulations including the illocutionary point indicating devices *pardon* and *sorry*.

References

Haselow, Alexander (2024), Politeness, speech acts and socio-cultural change. The expression of gratitude in the history of English. *Journal of Historical Pragmatics,* 25(3), 419–449. https://doi.org/10.1075/jhp.21005.has

Jucker, Andreas H. (2019), Speech act attenuation in the history of English: The case of apologies, *Glossa: A Journal of General Linguistics* 4(1), 45, 1–25. DOI: https://doi.org/10.5334/gjgl.878

Jucker, Andreas H., and Irma Taavitsainen (2008), Apologies in the history of English: Routinized and lexicalized expressions of responsibility and regret, in Andreas H. Jucker and Irma Taavitsainen (eds), *Speech Acts in the History of English.* (Pragmatics & Beyond New Series 176). Amsterdam/Philadelphia: John Benjamins, 229–244.

Williams, Graham (2018), Sincerity in Medieval English Language and Literature, London: Palgrave Macmillan.

Grumbling through time: The metadiscourse of complaints in historical letters and emails

Sofia Rüdiger and Rachele De Felice
(FU Berlin, The Open University)

Complaints are ever-present in our interactions, be it when making small talk about the weather (which is too cold), chatting about politics (which we disagree with), or engaging with the waiter in a restaurant about our food (which was too salty). These examples illustrate two main complaint types: 1) other-addressed complaints (the first two cases), which, from a present-day perspective have been shown to be essential in building rapport (Boxer 1993) and eliciting 'emotional reciprocity' (Günthner 1997), while being socially-stigmatized at the same time (e.g., Heinemann & Traverso 2009: 2381); and 2) complainee-addressed complaints (the third case), which are considered inherently face-threatening to the hearer (Olshtain & Weinbach 1987: 196). The complex interaction between this speech act and both own- and other-face considerations makes it an interesting case study from the perspective of metadiscourse (i.e., displays of reflective awareness (Haugh 2018) as performed in people's speech or writing (cf. Jucker 2020)), as speakers navigate the balance between explicitly expressing their stance and maintaining interpersonal harmony. In this article, we take a historical corpus linguistic perspective on complaint metadiscourse, to shed light on the variation of this function at different points in time.

We do this by investigating three datasets: 1) American English letters from the Late Modern English period (a subset of the *Corpus of Early American Literature*, CEAL; Höglund & Syrjänen 2016), 2) British English letters from the Late Modern English period (*British Telecom Correspondence Corpus*; Morton & Nesi 2020), and 3) present-day American English emails (*Clinton Email Corpus*; De Felice & Garretson 2018). We search these corpora for lexical items (and related lemmas and word forms) which were identified as salient in previous research on complaint metadiscourse in personal historical letters (Rüdiger in prep), for example, *complain*, *dissatisfaction*, *grudge*, *grumble*, *murmur*, and *remonstrate*. Based on the patterns of use of these terms, we answer the following research questions: 1) How are these lexical items used - are they IFIDs or do they perform other functions? 2) Are there any differences between the two historical varieties (i.e., British English and American English)? 3) How does this compare to data from the 20th/early 21st century? The study thus allows us to observe variation in the use of these metadiscursive items over time and space. In addition, the article makes methodological contributions by employing a number of under-used corpus resources and by attempting a first comparison of complaint-metadiscourse across the genres of handwritten and electronic correspondence from different time periods.

References

Boxer, Diana (1993), Social Distance and Speech Behavior: The Case of Indirect Complaints, *Journal of Pragmatics,* 19, 103–125.

De Felice, Rachele and Gregory Garretson (2018), Politeness at Work in the Clinton Email Corpus: A First Look at the Effects of Status and Gender, *Corpus Pragmatics,* 2, 221–242.

Günthner, Susanne (1997), Complaint Stories: Constructing Emotional Reciprocity among Women, in Helga Kotthoff and Ruth Wodak (eds), *Communicating Gender in Context*, Amsterdam: John Benjamins, 179–218.

Haugh, Michael (2018), Corpus-Based Metapragmatics, in Andreas H. Jucker, Klaus P. Schneider and Wolfram Bublitz (eds), *Methods in Pragmatics,* Berlin: De Gruyter, 619–644.

Heinemann, Trine and Véronique Traverso (2009), Complaining in Interaction, *Journal of Pragmatics,* 41(12), 2381–2384.

Höglund, Mikko and Kaj Syrjänen (2016), Corpus of Early American Literature, *ICAME Journal* 40, 17–38.

Jucker, Andreas H. (2020), *Politeness in the History of English. From the Middle Ages to the Present Day*, Cambridge: Cambridge University Press.

Morton, Ralph and Hilary Nesi (2020), Institutional Collaboration in the Creation of Digital Linguistic Resources: The Case of the *British Telecom Correspondence Corpus*, in Simon Popple, Andrew

Prescott and Daniel H. Mutibwa (eds), *Communities, Archives and New Collaborative Practices,* Bristol: Policy Press, 153–164.

Olshtain, Elite and Liora Weinbach (1987), Complaints: A Study of Speech Act Behavior among Native and Nonnative Speakers of Hebrew, in Jef Verschueren and Marcella Bertuccelli-Papi (eds), *The Pragmatic Perspective: Selected Papers from the 1985 International Pragmatics Conference,* Amsterdam: John Benjamins, 195–209.

Rüdiger, Sofia (in prep), *Historical Perspectives on the Speech Act of Other-Addressed Complaint – Other-Addressed Complaint Realization, Strategies, and Responses*.

Letter-writing and speech act analysis: Tracing (im)politeness patterns in historical Irish English

David Sotoca Fernández and Carolina Amador-Moreno
(University of Extremadura, University of Bergen)

This paper presents an overview of three different case studies focusing on the use of (im)politeness (Brown & Levinson 1987; Haugh & Culpeper 2018) strategies pertaining to the encoding of requests, apologies and reproaches in historical Irish English. It resorts to a subsection of CORIECOR (Corpus of Irish English Correspondence) (Amador-Moreno 2021) comprising missives exchanged by Irish emigrants that moved to the US and their relatives and close ones from 1700 to 1940. This subsection contains a total of 596 letters that had been annotated manually in order to extract these speech acts systematically. All these investigations focus on intimate discourse as described by Clancy (2005) and use different theoretical frameworks (Archer 2017; Blum-Kulka 1984) within (im)politeness studies to tackle the data at hand. The first case study deals with the use of the mental verb "hope" in three speech acts and observes its historical value as a mitigator among all of them. The case study focuses on reproaches, a speech act that has been so far overlooked and categorized within the notion of critique. The analysis shows the existence of two different kinds of reproaches in the data employing Corpus Linguistics tools to shed light on the relevant lemmas that play a key function in the encoding of either of them. The third case study takes a more traditional approach to the analysis of requests within the data. It uses a combination of Blum-Kulka's (1984) and Ackermann's (2023) taxonomy for the analysis of requestive speech acts and categorizes each instance to determine the level of directness and indirectness appearing in this subcorpus when performing this speech act.

The focus of the paper will be on methodological issues regarding the analysis of these phenomena in historical data as illustrated in the three case studies. It concentrates on the value of their results for Irish English as a linguistic variety as well as their limitations and the caveats they present for further research. The paper discusses their potential comparability with previous studies using the same set of data and/or similar corpora, focusing on the value of these pieces for the area of (im)politeness studies and their validity within this field.

References

Ackermann, Tanja (2023), Mitigating strategies and politeness in German requests, *Journal of Politeness Research,* 19(2), 355–389.

Amador-Moreno, Carolina P. (2022), Contact, Variation and Change: Mapping the History of Irish English through CORIECOR, *Nexus,* 2, 49–53.

Blum-Kulka, Shoshana and Elite Olhstain (1984), Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP), *Applied Linguistics*, 5(3), 196–213.

Brown, Penelope and Stephen C. Levinson (1987), *Politeness: Some Universals in Language Usage*, Cambridge: Cambridge University Press.

Clancy, Brian (2015), *Investigating Intimate Discourse: Exploring the spoken interactions of families, couples and friends,* Routledge Taylor and Francis Group.

Haugh, Michael and Jonathan Culpeper (2018), Integrative Pragmatics and (Im)politeness Theory, in Cornelia Ilie and Neal R. Norrick (eds), *Pragmatics and its Interfaces,* John Benjamins, 213–239.

- *Hello, are you kidding?* A study of speech acts realized by the discourse markers *hello* in English and *hallo* in Norwegian

Jan Svennevig and Ingrid Kristine Hasund
(University of Agder)

This paper explores the methodological challenges involved in a cross-linguistic corpus-based analysis of the discourse markers *hello* in English and *hallo* in Norwegian. *Hello/hallo* are interesting in a speech act perspective as they can form an utterance alone and perform a range of different speech acts depending on the context. While the traditional vocative uses as a greeting and a summons (Schegloff, 1968) are well established, their evolving non-vocative usages—such as expressing reproach or surprise—remain underexplored (Andersen, 2014). To the best of our knowledge, there exists no in-depth study of *hello* as a discourse marker in English, and Svennevig (2012) is to date the only study of *hallo* in Norwegian. Using data from three corpora of informal spoken Norwegian with audio-linked transcripts—the UNO corpus of teenage language (1997-1998), the Big Brother corpus of young adult language (2001) and the NoTa Oslo corpus (different age groups, 2005) —Svennevig found that *hallo* may be used with three main non-vocative functions: as a reproach to an addressee for having said or done something inappropriate or incorrect, as a negative evaluation of some event, and as an announcement of a newsworthy or interesting event.

To investigate potential cross-linguistic pragmatic borrowing from English to Norwegian, we initially examined several English corpora from around the same time period as the Norwegian data: the CABNC corpus of informal conversations (1980s-1990s), the CallHome and CallFriend corpora of unscripted telephone conversations (1997) and the Santa Barbara Corpus of Spoken American English (1990-1997). Surprisingly, few instances of non-vocative *hello* were found within these datasets. While the reproach and announcement functions appeared, the expected negative evaluation function was absent. This led us to expand our methodological approach. We complemented the corpus searches with a diachronic survey of English dictionaries (1964-2024), revealing that only the reproach function aligns with Norwegian *hallo*. Contrary to the Norwegian data, the English dictionaries frequently list *hello* as 'expressing surprise'; without context, however, it is hard to interpret this speech act.

Recognizing the importance of non-verbal cues—such as gaze and gesture—in interpreting *hello/hallo*, we incorporated multimodal data. Analysis of the US TV series *Seinfeld* scripts (1989-1998) uncovered a few instances of reproach and negative evaluation, but no cases of announcements. Further, using a subset from 2016 of the NewsScape corpus of US TV news, which integrates video with transcripts, we identified approximately 80 cases of non-vocative *hello*. In addition to the same three functions described by Svennevig (2012), we observed *hello* used with three other functions not previously documented in research: as an epistemic marker for self-evident truths, as positive evaluation, and as a metaphorical summons.

This study underscores the complexities of using older data for diachronic analysis and highlights the necessity for multimodal resources to capture the full communicative spectrum of discourse markers. Further, it discusses issues relating to metadata, annotation and sound/transcript quality and alignment across different corpora. Aligning with the workshop's focus on methodological rigor, we contribute to broader discussions on dataset comparability and corpus-based speech act research.

References

Andersen, Gisle (2014), Pragmatic borrowing, *Journal of Pragmatics, 67*, 17–33. http://dx.doi.org/10.1016/j.pragma.2014.03.005

Schegloff, Emanuel A. (1968), Sequencing in conversational openings, *American Anthropologist,* 70, 1075–1095.

Svennevig, Jan (2012), "og jeg bare hallo liksom" Diskursmarkøren hallo i samtale, *Språk och interaktion* 3, 157–174. http://hdl.handle.net/10138/37523

Workshop 2: Conventionalisation, specialisation, institutionalisation: Exploring communicative practices and genre developments in letter writing

Convenors:
Lisa Lehnen (JMU Würzburg)
Theresa Neumaier (TU Dortmund)
Ninja Schulz (JMU Würzburg)

The delineation of genres and subgenres is a challenging task for the compilation of corpora that aim at comparability across varieties and/or across time as genres constantly evolve in accordance with socio-historical changes. At the same time, their variability is limited by the fact that they also have to remain recognisable for language users (Brinton 2023: 186). Understanding the factors and forces shaping genres within specific socio-cultural settings is thus a key element in designing corpora and interpreting synchronic and diachronic variation. Correspondence is particularly interesting in this respect as letters are usually a) non-edited, b) interactive and c) sensitive to technological progress.

With respect to a), written genres are often regulated by institutions or public players, regarding mostly the publication process and sometimes also the production process. This is not necessarily the case for correspondence, where a less clearly defined group of individuals produces texts and engages in establishing norms by relying on conventions of language use that they are familiar with (Claridge 2017: 186). However, not all language users have access to the same models of correspondence and conventions for new sub-genres emerging from changing communicative needs (such as company-internal business correspondence) have to be established first. Regarding b), the socio-pragmatic function of letters (as a form of written interaction) (Bergs 2007) makes them especially sensitive to cultural norms regarding, for instance, politeness, stance-taking, and expressions of deference, which will affect the conventions emerging in different cultural and socio-political settings (including modifications made in cross-cultural communication, for instance in international business correspondence). Thus, while politeness and communicative conventions have some universal characteristics, they are largely dependent on context. The resulting variation is induced by culture-specific understandings of face, rights and obligations and interactional goals (Spencer-Oatey 2008). In addition, norms and conventions have changed through time due to increasing language contact and economic and societal transformations, e.g. the spread of literacy, democratisation, or changing gender roles (Bruns & Kranich 2021; Jucker 2020; Loureiro-Porto 2021). Regarding c), technological developments have changed letter writing considerably over the centuries, including changes in the mode of production (handwritten, typewritten, electronic, etc.), means of transport (railway, steam ships, etc.), and services (e.g. penny post, internet), all leading to an increasingly reduced time lag between sending and receiving letters (from several months to instant communication) and the expansion of the group of people participating in the practice of producing correspondence.

In this workshop, we aim at bringing together researchers who explore correspondence from a diachronic perspective. The contributions cover the time span from Early Modern English to the 20th century exploring diverse contexts, such as health communication, threatening letters, business correspondence and pauper letters from Britain, the US and Hong Kong. Key issues include but are not limited to:

- Conventionalisation: What language practices have become typical in correspondence over time? To what extent are these sensitive to cultural setting, politeness norms, and influences from other media etc.?
- Specialisation: What (sub)genres have emerged (for instance business letters, threatening letters, etc.)? To what extent has the genre diversified? What influences between specialised subgenres and beyond genre boundaries can be identified?
- Institutionalisation: What role does correspondence have in different domains? Which practices have become part of institutional discourses? Which groups have been included in and excluded from these practices?

By investigating historical correspondence from and across English-speaking communities and contexts, we want to initiate a discussion about new approaches to tracing and theorising genre developments and the complexities involved therein.

References

Bergs, Alexander (2007), Letters: A new approach to text typology, in Terttu Nevalainen & Sanna-Kaisa Tanskanen (eds), *Letter Writing*, Amsterdam: John Benjamins Publishing Company, 27–46.

Brinton, Laurel J. (2023), *Pragmatics in the History of English*, Cambridge: Cambridge University Press.

Bruns, Hanna and Kranich, Svenja (2021), Terms of Address: A Contrastive Investigation of Ongoing Changes in British, American and Indian English and in German, *Contrastive Pragmatics,* 3(1), 112–143. https://doi.org/10.1163/26660393-BJA10025

Claridge, Claudia (2017), Discourse-based approaches, in Laurel J. Brinton (ed), *English Historical Linguistics: Approaches and Perspectives*, Cambridge: Cambridge University Press, 185–217.

Jucker, Andreas H. (2020), *Politeness in the History of English: From the Middle Ages to the Present Day,* Cambridge: Cambridge University Press.

Loureiro-Porto, Lucía (2021), Linguistic Colloquialisation, Democratisation and Gender in Asian Englishes, in Tobias Bernaisch (ed), *Gender in World Englishes*. Cambridge: Cambridge University Press, 176–204.

Spencer-Oatey, Helen (2008), Face, (Im)politeness and Rapport, in Helen Spencer-Oatey (ed), *Culturally Speaking: Culture, Communication and Politeness Theory* (2nd ed), London: Continuum, 11–47.

Understanding pragmalinguistic choices in appellative letters – focus on early 20th-century business and extortion letters

Theresa Neumaier, Ninja Schulz and Lisa Lehnen
(TU Dortmund, JMU Würzburg)

Correspondence provides a rich data source for studies in historical sociolinguistics, genre analysis, language variation and change. However, to fully exploit the potential of the data, it is important to be aware of external factors relevant for their production and acknowledge the options and limitations for deriving such metainformation. Socio-historical, cultural and technical aspects, e.g. the socio-demographic background of the writer, societal structure, communication purpose, and the production circumstances, need to be understood to enable a comprehensive linguistic analysis and valid interpretation. Although specific grammatical features or discourse-pragmatic units, e.g. modal verbs and requests, can be easily located and formally analysed in different types of datasets, interpreting their use and function requires a clear conception of their embedding in the respective context. Similarities and differences in the occurrence of these features may be inflicted by conventions established within the (sub)genre at large but also on the level of the speech community, specific social networks, or even the individual.

To illustrate the relevance of contextualisation, we investigate two sets of early 20th-century letters, written in different cultural settings and with varying communicative purposes. The first dataset consists of business correspondence from Hong Kong collected from archives in London and Cambridge, the second of extortion letters collected from the Metropolitan Police's Threatening Letter Book. We focus on directives, which are well-researched in cross-cultural comparisons and omnipresent in business-like correspondence. Our letters are appellative by nature and similar in their communicative intent - the sender asks for some action from the addressee, thus exerting the illocutionary force of a directive (Bergs, 2007, p. 33). However, the ways the letter writers set out to achieve their goals differ considerably as the business letters include socially sanctioned directives, while in extortion letters the directive is realised as an illicit speech act. We show that factors traditionally used to describe the seriousness of face-threatening acts like requests (social distance, power, rank of imposition) cannot fully explain pragmalinguistic choices. Regarding power, for instance, some of the writers in our corpus of Hong Kong business correspondence have inherent status (Bargiela-Chiappini et al., 1996, p. 637) in Hong Kong due to their societal activities beyond the business

context, while others only have relative status within the company. In extortion letters sent to high-status recipients (e.g. the Prime Minister), the (mostly anonymous) writers use intertextual references to depict themselves as members of powerful groups, thus reducing the social distance between themselves and their recipient and adding force to their demands.

For both datasets, the specialisation of letter types leads to a conventionalisation of realisation strategies and the development of formulae, which can even override politeness norms once the transactional function of the letter is established. On a theoretical level, our findings therefore imply that we need to find ways to conceptualise sociopragmatic factors in a more fine-grained manner, allow for interactions between them and integrate dynamic processes of genre-internal conventionalisation and specialisation to understand pragmalinguistic choices in correspondence.

References

Bargiela-Chiappini, Francesca and Sandra J. Harris (1996), Requests and status in business correspondence, *Journal of Pragmatics,* 26(5), 635-662. https://doi.org/10.1016/0378-2166(96)89191-0

Bergs, Alexander (2007), Letters: A new approach to text typology, in Terttu Nevalainen and Sanna-Kaisa Tanskanen (eds), *Letter writing,* John Benjamins Publishing Company, 27–46.

Between culture-specific practices and general trends in letter writing: Setting the scene for tracing conventionalisation, specialisation and institutionalisation

Ninja Schulz, Lisa Lehnen and Theresa Neumaier
(JMU Würzburg, TU Dortmund)

With corpora, linguists seek to provide a sample of language use representative of a specific variety, genre, time period or topic. Depending on the breadth and aim of the corpus project, parallel corpora have been compiled for comparisons of language use across time and varieties. Correspondence has traditionally been included in (diachronic) corpora since it approaches spoken language and is thereby more inclusive of different (literate) social strata. From a superordinate perspective, correspondence can be clearly delineated as a genre, being marked by specific situational characteristics (Biber & Conrad 2009: 40). However, it also shows a high degree of genre-internal heterogeneity. Letters are extremely culturally-sensitive and dynamic regarding who has access to writing/reading them, how they are written and why. Over time, different types of letters evolve based on the communicative purpose, with some letter writing practices becoming established and others disappearing (e.g. the emergence of pauper letters in consequence of the Old Poor Law). On the other hand, models for some letter types are not always available in society, especially for those which are not socially sanctioned, i.e. "illicit genres" (Bojsen-Møller et al. 2020). Therefore, to identify factors affecting specific linguistic choices within the genre, the respective sociocultural context must be considered. The variation in the genre of correspondence across time and culture thus problematises corpus compilation. Nevertheless, letters constitute very rich datasets since even gaps in the data (e.g. regarding social variables like gender) can be taken as indicative of the sociocultural context.

In our introductory talk, we showcase the processes of conventionalisation, specialisation, institutionalisation in correspondence on the basis of three contexts involving different periods of time, cultural settings and interactional goals: 19th-century letter-writing guidelines in Madeira, threatening communication in Late Modern English, and 20th-century business correspondence from Hong Kong. In terms of conventionalisation, we trace the development of communicative practices that are nowadays common in letter writing but had not been established in previous periods. We show that the structure of letters has undergone distinct changes over time, with elements newly emerging, disappearing, or becoming obligatory (e.g. the salutation in extortion letters). Furthermore, we analyse changes on the pragmalinguistic level in speech act realisation, especially levels of directness in requests and threats. Concomitant with conventionalisation, our data shows that free variation in the linguistic choices decreases

or becomes restricted to specific types of letters. Changes in the sociopragmatic functions of letter types (e.g. the disappearance of gossip from business correspondence) can thus be taken as indicative of processes of specialisation. Such an evolution is reflective of institutional changes relating to company structure, technological advances, or legal contexts.

Considering that the variables influencing language use in correspondence are highly culture- and context-dependent, they must be identified to enable comparisons across time and varieties. While our presentation is necessarily selective, it sets a departure point for the discussion of genre developments in letter writing and the challenges and opportunities these pose to corpus linguistics.

References

Biber, Douglas and Susan Conrad (2009), *Register, genre, and style*, Cambridge: Cambridge University Press.

Bojsen-Møller, Marie, Sune Auken, Amy Devitt and Tanya Christensen (2020), Illicit Genres: The Case of Threatening Communications, *Sakprosa,* 12(1), 1–53.

Conventionalization of politeness strategies in Hong Kong business correspondence: A diachronic perspective on the use of requests

Carina Stick
(JMU Würzburg)

While business correspondence is generally perceived as a highly conventionalized genre, studies have shown that this is a rather recent development. Up to the 19th century, business letters were virtually the only way of communicating over long distances, which is why they often contained private news and gossip (Dossena 2006: 176). It is only from the mid-19th century that, with the development of firm structures from small family businesses to larger companies, a new business ethos focusing on transactional approaches emerged, ultimately leading to the conventionalization of the genre (Del Lungo Camiciotti 2006: 156).

This paper investigates the conventionalization of politeness strategies in early 20th century business correspondence from Hong Kong through an analysis of requests and their accompanying moves. While requests have been studied extensively since the 1980s (e.g. Blum-Kulka 1987, Del Lungo Camiciotti 2008), the fact that these often occur in combination with other moves, such as justifications (Kong 1998) or apologies (Blum-Kulka & Olshtain 1989), has been neglected to date. Therefore, this paper investigates the following questions:

- How has the use of requests and their accompanying moves in Hong Kong business correspondence changed from the 1900s to the 1940s?
- Which expressions, request strategies or supportive moves have become conventionalized over time?

The study is based on a corpus of business letters sent from or to two Hong Kong based businesses, namely the holding company Jardine Matheson & Co. and the Hong Kong and Shanghai Banking Corporation (HSBC). Three sub-corpora, each covering a decade (the 1900s, 1920s and 1940s), have been compiled and digitized with every sub-corpus consisting of roughly 25.000 words.

The project relies on a mixed method approach: after manually identifying typical words and phrases that indicate a request by reading through the letters of a test corpus, using AntConc each subcorpus was searched for these expressions to find all requests. These were then coded and analyzed following an adapted coding scheme which is based on Blum-Kulka and Olshtain's (1989) coding manual, so that some initial results can be presented. It is expected that the results will, due to the reduction of social hierarchies and the establishment of corporate companies in the 20th century, show a less overly polite style and an increase in directness in the formulation of requests. Furthermore, it is expected that certain expressions

and request strategies have been conventionalized and are routinely employed based on the situational context rather than semantics.

References

Blum-Kulka, Shoshana (1987), Indirectness and politeness in requests: Same or different? *Journal of Pragmatics,* 11(2), 131–146.

Blum-Kulka, Shoshana and Olshtain, Elite (1989), *Cross-cultural pragmatics: Requests and apologies,* Norwoord, NJ: Ablex.

Dossena, Marina (2006), Forms of self-representation in nineteenth-century business letters, in Marina Dossena and Irma Taavitsen (eds), *Diachronic perspectives on domain-specific English,* Bern: Lang, 173–190.

Del Lungo Camiciotti, Gabriella (2006), "Conduct yourself towards all persons on every occasion with civility and in a wise and prudent manner; this will render you esteemed": Stance features in nineteenth century business letters, in Marina Dossena & Susan M. Fitzmaurice (eds), *Business and official correspondence: Historical investigations*, Bern: Lang, 175–192.

Del Lungo Camiciotti, Gabriella (2008), Two polite speech acts from a diachronic perspective: Aspects of the realisation of requesting and undertaking commitments in the nineteenth century commercial community, in Andreas H. Jucker and Irma Taavitsainen (eds), *Speech acts in the history of English,* Amsterdam: John Benjamins, 115–131.

Kong, Kenneth C. C. (1998), Are simple business request letters really simple? A comparison of Chinese and English business request letters, *Text & Talk*, 18(1), 103–141.

"I don't comprehend this intercourse of seal'd letters" – Contractions in the Mary Hamilton Papers

Monique Tschachtli
(University of Zurich)

This paper investigates the use of contractions in Mary Hamilton's private letters and diaries, alongside the letters of her correspondents. Contractions in eighteenth-century private letters were considered informal and vulgar following criticism from Swift and Addison (Haugland 1995) and can therefore be interpreted as markers of familiarity. Previous research on contractions in Elizabeth Montagu's correspondence finds that they are most frequent in letters by female writers, family members, social equals, and older writers and that Montagu's use increased over time despite mounting criticism of the form (Sairio 2009, 2010). Hamilton's writings, recently made accessible in an electronic corpus, have not been investigated thoroughly. Moreover, previous research on contraction use in eighteenth-century letters has relied on descriptive statistics, only. This paper addresses a research gap in the field of historical sociolinguistics by examining contraction use with a multi-factorial approach and utilising a novel dataset.

The paper focuses on 'participle contractions' (e.g. *lov'd*, *lik'd*, *receiv'd*) and 'negation contractions' (e.g. *don't*, *can't*, *couldn't*), contrasting them with their full forms. The study answers the following research questions: How does Hamilton's contraction use vary according to contraction type, over time, between her letters and diaries, and based on her addressee's age, gender, social rank, and relationship to her? How does her correspondents' contraction use vary according to contraction type, over time, and based on their age, gender, social rank, and relationship to her? Which variables most strongly affect writers' contraction use and how do they interact? By examining these questions, the study explores intra- vs. inter-speaker differences within the broader context of the conventionalisation of a stigmatised feature in the specialised genre of personal letters.

The data are extracted from *Unlocking the Mary Hamilton Papers* (Barker et al. 2019) on *CQPweb* (Hardie 2022), including full and contracted participle and negation from Hamilton's diaries, out-letters, and in-letters. Variables include author and recipient gender, age (relative to Hamilton), social rank,

relationship (to Hamilton), decade, and period (of Hamilton's life). The data are analysed using Random Forest models, combining the 'ctree' and 'cforest' functions in R.

The results reveal that Hamilton uses participle contraction frequently, but categorically avoids negation contraction. She uses participle contraction more frequently in her diaries than in her letters, and her contraction use decreases significantly over time. This pattern varies based on her correspondent's age, social rank, and relationship to her. In her correspondents' letters, participle contraction is likewise more frequent than negation contraction, though their contraction use shows a less marked decrease over time. This pattern varies based on their gender, age, social rank, and relationship to Hamilton. The Random Forest analyses reveal the complex interactions between these variables in shaping contraction use, challenging especially the importance ascribed to gender by previous studies (Sairio 2009, 2010, 2018).

Thus, this paper suggests that the stigmatisation of contractions limited their perceived appropriateness to specific contexts. This provides new insights into the conventionalisation of familiar features in personal letters as a specialised genre.

Sources

Barker, Hannah, Sophie Coulombeau, David Denison, Tino Oudesluijs, Cassandra Ulph, Christine Wallis and Nuria Yáñez-Bouza (2019), *Unlocking the Mary Hamilton Papers*. https://www.maryhamiltonpapers.alc.manchester.ac.uk/ (5 September 2024).

Hardie, Andrew (2022), *The Mary Hamilton Papers*: Powered by CQPweb, Corpus, CQPweb. https://cqpweb.lancs.ac.uk/hamilton/ (13 September 2024).

References

Haugland, Kari E. (1995), Is't allow'd or ain't it? On Contraction in Early Grammars and Spelling Books, *Studia Neophilologica,* 67(2), 165–184. doi:10.1080/00393279508588159.

Sairio, Anni (2009), *Language and Letters of the Bluestocking Network: Sociolinguistic Issues in Eighteenth-Century Epistolary English*, Société néophilologique.

Sairio, Anni (2010), "if You think me obstinate I can't help it": Exploring the Epistolary Styles and Social Roles of Elizabeth Montagu and Sarah Scott, in Päivi Pahta, Minna Nevala, Arja Nurmi and Minna Palander-Collin (eds), *Social Roles and Language Practices in Late Modern English*, Amsterdam: John Benjamins Publishing Company, 87–109. doi:10.1075/pbns.195.05sai.

Sairio, Anni (2018), Weights and Measures of Eighteenth-Century Language: A Sociolinguistic Account of Montagu's Correspondence, *Huntington Library Quarterly,* 81(4), 633–656. doi:10.1353/hlq.2018.0024.

Conventionalisation of health communication in Early Modern English letters

Oriana Yim
(Heidelberg University)

Everyone talks about health and sickness, not just to doctors, but also to family members, friends, and acquaintances. So how do people talk about health and what pragmatic functions does this communication perform? Most historical pragmatic research regarding health and illness concerns itself with medical discourse and texts (for an exception, see Taavitsainen 2011). However, this does not make up the entirety of health communication. In contrast to existing research on historical medical discourse, this presentation is concerned with the forms and functions of laypeople's health communication. Naturally anyone can partake in informal health communication; and I argue that everyone does communicate health. Health, as a topic of informal conversation, is reasonably diachronically stable – a *tertium comparationis* – this makes informal health communication an almost universal phenomenon, and a valuable subject for (diachronic) analysis.

This project analyses laypeople's informal health communication in Early Modern epistolary letters, as letters offer an excellent channel to examine laypeople's language since they are authentic, neither fictionalised, nor reported. Following Fitzmaurice (2002) and Palander-Collins (2002), I take a historical pragmatic approach to investigating epistolary data and am concerned with how health communication performed pragmatic functions which were conventionalised by language users.

My data comes from the *Parsed Corpus of Early English Correspondence* (PCEEC). I present a case study on the Arundel collection of 78 letters (1589–1680). This collection was selected because it is small enough to locate health communication instances through close reading (as proposed by Kohnen 2007) and pragmatically annotate them, yet large enough to include many instances. This collection is from a prominent family in English history, therefore there is contextualising information available for many letters. This analyse is a pilot study in anticipation of an analysis of the whole corpus. Following health communication categorisation, I qualitatively analysed occurrences, interpreting pragmatic meanings including speech acts and politeness, and considering various layers of context.

Preliminary findings show health communication, particularly sharing and requesting health reports, was a conventionalised means to strengthen and maintain relationships and keep communication channels open (Locher and Graham 2010). Health reports were also employed as justifications to mitigate against impoliteness. Additionally, these findings support Fitzmaurice's (2002) analysis of advice in medical council, as they show that giving health advice was a pragmatic act to demonstrate empathy and affection. Conventionalisation of health communication is most visible in letter-concluding health formulae constructed as undefined well-wishes. Comparing this data to informal PDE, processes of pragmaticalisation and speech act attenuation are perceived (Jucker 2019), particularly in well-wishes and reference to God which are no longer as sincere. To conclude, informal health communication is a device worthy of analysis because it offers a concise, yet widespread and stable, focus to analyse conventionalised interpersonal interaction; by concentrating on such a device, rather than confining analysis to one speech act, researchers build multifaceted perception of interactions. Through the study of this linguistic device this research contributes to our understanding of relational functions of language.

Data

Parsed Corpus of Early English Correspondence 2, parsed version (2022), revised and corrected by Beatrice Santorini, annotated by Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen, compiled by the CEEC Project Team. https://github.com/beatrice57/pceec2.

References

Fitzmaurice, Susan M. (2002), *The Familiar Letter in Early Modern English: A Pragmatic Approach*, Pragmatics & beyond New Series, Amsterdam / Philadelphia: John Benjamins Publishing Company.

Jucker, Andreas H. (2019), Speech Act Attenuation in the History of English: The Case of Apologies, *Glossa: A Journal of General Linguistics,* 4(1), 1–25. https://doi.org/10.5334/gjgl.878.

Kohnen, Thomas (2007), Text Types and the Methodology of Diachronic Speech Act Analysis, in Susan Fitzmaurice and Irma Taavitsainen (eds), *Methods in Historical Pragmatics*, Berlin, New York: Mouton de Gruyter, 139–66.

Locher, Miriam A. and Sage L. Graham (2010), *Interpersonal Pragmatics*, Berlin / New York: De Gruyter Mouton. https://doi.org/10.1515/9783110214338.

Palander-Collin, Minna (2002), Tracing Patterns of Interaction in Historical Data, in Helena Raumolin-Brunberg, Minna Nevala, Arja Nurmi, and Matti Rissanen (eds), *Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen*, Mémoires de La Société Néophilologique de Helsinki 61, Helsinki: Société Néophilologique, 117–34.

Taavitsainen, Irma (2011), Dissemination and Appropriation of Medical Knowledge: Humoral Theory in Early Modern English Medical Writing and Lay Texts, in Irma Taavitsainen and Päivi Pahta (eds), *Medical Writing in Early Modern English*, Studies in English Language. Cambridge: Cambridge University Press, 73–93. https://doi.org/10.1017/CBO9780511921193.

Taavitsainen, Irma, and Andreas H. Jucker (2010), Trends and Developments in Historical Pragmatics, in Andreas H. Jucker and Irma Taavitsainen (eds), *Historical Pragmatics*, Handbooks of Pragmatics 8, Berlin / New York: De Gruyter Mouton, 3–29.

Conventionalisation, specialisation, institutionalisation: Discussing communicative practices and genre developments in letter writing

Ninja Schulz, Theresa Neumaier and Lisa Lehnen
(JMU Würzburg, TU Dortmund)

During the closing session, the workshop convenors will synthesise the main themes and conclusions arising from the presentations. There will be a particular focus on discussing questions and issues raised in the workshop proposal relating to the conventionalisation of communicative practices, the specification of correspondence into letter types, the institutionalisation of letter writing, and the variation that can be observed across time, socio-historical contexts, and cultural settings. We will also talk about options for further collaboration and a joint publication.

Workshop 3: Per studia contrastiva ad astra: Looking back, leaping forward

Convenors:
Lot Brems (University of Liège)
Lobke Ghesquière (University of Mons & KU Leuven)
Gudrun Vanderbauwhede (University of Mons)

Contrastive linguistics now has a rich history and this workshop is conceived as an opportunity to bridge insights from earlier research with novel research in the field. It aims to offer a platform to revisit key studies and frameworks from the past, understand their impact on current scholarship, and explore how research today can build upon or reinterpret foundational work. As such, this workshop wants to create dialogue between "then" and "now".

The objectives of the workshop and the papers it offers a forum to include

1. **revisiting seminal work** by encouraging scholars to re-engage with influential works and to discuss their relevance to contrastive linguistics in today's research landscape;
2. **expanding the scope** of existing studies by applying new methods, using new corpora, involving more languages and/or re-investigating the data or issues from new angles (e.g. synchrony vs. diachrony; pedagogical, theoretical or applied perspectives);
3. **exploring the possible enrichment** of contrastive linguistics studies through intersecting fields such as pragmatics, translation studies, second language acquisition (SLA), cognitive or computational linguistics.

The enrichment by/of contrastive linguistics of/by neighbouring fields is explored in many of the submitted contributions. In her study on requests in English and Swedish, Aijmer explores how contrastive studies can enrich the field of pragmatics, for example. Conversely, a number of other submissions focus on how the field of contrastive linguistics can be enriched by a.o. translation studies, construction grammar and diachronic approaches.

The abstracts submitted by Ebeling & Ebeling, Hasselgård and Egan all explore how insights from corpus-based translation studies can enrich previous findings from contrastive linguistics. Ebeling & Ebeling focus on cognate noun/verb pairs in English and Norwegian, expanding the study to more language pairs than ever before and including bidirectional parallel corpus data. Hasselgård will revisit and build on her own and others' work on habitual verbal expressions in English and Norwegian and Egan will expand earlier findings from contrastive studies on posture verbs with new insights from parallel English-Norwegian data.

Through a study of argument structure constructions in English and German, Uhrig & Herbst will show how new corpora and a construction grammar approach allow us to better and more fully understand the constructions under scrutiny, both language-specifically and cross-linguistically. The results of previous contrastive findings on noun phrase density will be refined by insights from Krielke's diachronic study of noun phrase density in English and German in academic writing.

Also focusing on academic writing, Rørvik will revisit and partly replicate seminal work by Fløttum et al. (2006) on first person pronouns by studying the topic in recent Norwegian and English data, thus enabling a diachronic comparison. This study, like the studies by Krielke and Uhrig & Herbst, has pedagogical goals, aiming to contribute to textbook design and the development of (foreign language) teaching materials.

Like Rørvik and Hasselgård, Levin et al. will also revisit earlier work and expand on previous work by using new corpus data from the trilingual Linnaeus University English-German-Swedish corpus, allowing them to refine and adjust earlier contrastive findings. Similarly, Malá's study of noun phrase complexity in in English and Czech children's literature is possible only through new corpus developments, as her research draws on data from the parallel corpus InterCorp, whose most recent editions have been annotated using the universal dependencies framework (Rosen 2023) and includes several measures of syntactic complexity and lexical diversity.

Clearly, the submitted topics cover a diverse range of linguistic phenomena and languages, but they all share the aim of showing the merits of existing studies while showcasing how novel insights, techniques and approaches can enrich past findings.

Using contrastive corpora for investigating speech acts from an English-Swedish perspective - the case of requests

Karin Aijmer
(University of Gothenburg)

A major issue in pragmatics is to identify the linguistic manifestations of speech acts such as requesting, offering, suggesting, etc. It is also a topic to which corpora and corpus linguistics can contribute. Corpus linguists have been successful in investigating the functions of speech acts having a conventionalized form (cf. e.g. Deutschmann 2003 on apologies). However, other methodologies are needed to identify all (or a representative sample of) the patterns associated with a speech act function. A promising trend in monolingual corpus-based speech act studies has been to proceed from the definition of a speech act function to identify how the speech act is realized by manually reading through the text or by utilizing the results from previous research (see e.g. Pöldvere et al. 2022). The aim of my presentation is to extend the function-to-form methodology to the contrastive investigation of speech acts. The case study is an investigation of how polite requests are performed in English and Swedish using the English-Swedish Parallel Corpus as data for the analysis (Altenberg and Aijmer 2000).

Requests have been classified by Searle (1976) as directives having certain properties or felicity conditions which must be fulfilled by the patterns categorized in this way. The hearer must, for example be able and willing to carry out an action. It would be difficult to search for all the requestive forms manually in order to investigate their correspondences in the other language. The cross-cultural common core can therefore be taken to be interrogatives or declaratives with a modal auxiliary (and usually a second person subject). I therefore began to search for the occurrences of these patterns in the corpus in the English and Swedish original texts. The English search items are, for instance, *can you, could you, will you, you could, you would, etc.* but also *please, maybe, I think, just* which can be supposed to collocate with a polite request in their mitigating functions. I also searched for patterns with *please* (or one of the other markers) manually excluding examples where it was not followed by a request. Specifically, the Swedish correspondences of utterances containing *please* as a politeness marker contained patterns associated with requesting (which might be unexpected since they differed from the patterns in the English originals) such as *jag måste be er* ('I must ask you'), *var vänlig (snäll, bussig, god) och* ('be kind and'), *ni kan väl ('*you can I suppose'), *det går bra att* (lit. 'it goes well that'), *du kanske skulle* ('you perhaps should´). The preliminary findings suggest that we can get a rich description of the realizations of the speech act of requesting in the compared languages highlighting similarities and differences between the languages by using the corpus both as a comparable corpus and as a translation corpus.

References

Altenberg, Bengt and Karin Aijmer (2000), The English-Swedish Parallel Corpus. A resource for contrastive research and translation studies, in Christian Mair and Marianne Hundt (eds), *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20),* Freiburg im Breisgau 1999, Amsterdam/Atlanta: Rodopi, 15–33.

Deutschmann, Mats (2003), *Apologising in British English,* Institutionen för moderna språk, University of Umeå.

Pöldvere, Nele, Rachele De Felice, and Carita Paradis (2022), *Advice in conversation: Corpus pragmatics meets mixed methods*, Cambridge University Press.

Searle, John R. (1976), A classification of illocutionary acts, *Language in Society* 5(1), 1–23.

Issues of comparability and sameness in contrastive critical discourse studies

Niall Curry
(Manchester Metropolitan University)

Contemporary corpus-based contrastive linguistics is epitomised by concerns with comparability, data quality, and sameness. Theoretical notions of the *tertium comparationis*, equivalence, and sameness and identity govern how we validate comparisons in contrastive studies (Krzeszowski, 1984), questions of methodological directionality and convergent and divergent approaches influence research outcomes (Chesterman, 2007), and issues of alignment in data design, data quality, and representation (Johansson, 2003) pose challenges for corpus linguists engaging in contrastive work. Over the last 30 years, there has been an increased effort to merge contrastive and corpus methodologies (e.g., Hasselgård, 2020; Johansson, 2003; McEnery & Xiao, 2010) and unpack approaches to contrastive analysis in a range of subfields of corpus linguistics (e.g., corpus-assisted discourse studies, Vessey, 2013). However, in the wider literature, theoretical constructs in contrastive linguistics have had a relatively limited impact on multilingual and comparative research, while the use of corpus linguistics approaches appears to be growing across and beyond applied linguistics (Pérez-Paredes & Curry, 2024). In the context of critical discourse studies, for example, the use of corpus linguistics is largely normalised. Yet, in contrastive and multilingual critical discourse studies, the role of notions such as equivalence, sameness, and the *tertium comparationis* are largely unaddressed. As such, there is a need to critically assess the relevance of these fundamental concepts in critical discourse studies and delineate their effective operationalisation therein.

To operationalise these key concepts, in this talk I discuss two empirical studies, focusing on a contrastive analysis of 1) Brexit and 2) COVID-19 discourses in expert communication in English, French, and Spanish. For each analysis, I use themed corpora of academic news blog posts from *The Conversation*'s English, French, and Spanish language sites. *The Conversation* is an international site used by academics to disseminate their research to the public. With over 40 million monthly readers, of whom over 80% are non-academics, these texts are largely designed to inform the public of research developments across the spectrum of academic disciplines. These blogs posts have been found to demonstrate cultural differences across languages and reconstruct cultural and ideological perspectives (e.g., Curry, 2024). As such, they offer a valuable site for understanding the nature of expert communications, globally. For the analysis of Brexit discourses, I use a comparable corpus to conduct a convergent critical discourse analysis based on a comparison of keywords in each corpus. For the analysis of COVID-19 discourses, I use a parallel corpus of academic news blogs posts and their translations to conduct a divergent analysis, using transitivity analysis. Through these studies, I return to foundational concepts in contrastive linguistics to bolster corpus approaches to critical discourse studies. In so doing, I draw attention to the affordances of contrastive research and its theoretical underpinnings for offering a complex and layered understanding of globalised and localised social challenges.

References

Chesterman, Andrew (2007), Similarity Analysis and the Translation Profile, *Belgian Journal of Linguistics*, 21(1), 53–66.

Curry, Niall (2024), Questioning the climate crisis: A contrastive analysis of parascientific discourses, *Nordic Journal of English Studies*, 23(2), 235–267.

Hasselgård, Hilde (2020), Corpus-based contrastive studies: Beginnings, developments and directions. *Languages in Contrast*, 20(2), 184–208.

Johansson, Stig (2003), Contrastive linguistics and corpora, in Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson (eds), *Corpus-based approaches to contrastive linguistics and translation studies,* Brill Academic Pub, 31–44.

Krzeszowski, Tomasz P. (1984), Tertium comparationis, in Jacek Fisiak (ed), *Contrastive linguistics: Prospects and problems*, Berlin, New York: De Gruyter Mouton, 301–312.

McEnery, Tony and Richard Xiao (2010), *Corpus-based contrastive studies of English and Chinese*. Routledge.

Pérez-Paredes, Pascual and Niall Curry (2024), Epistemologies of corpus linguistics across disciplines, *Research Methods in Applied Linguistics*, 3(3), 100–141.

Vessey, Rachelle (2013), Challenges in cross-linguistic corpus-assisted discourse studies, *Corpora*, 8(1), 1–26.

From star (N) to star (V): A contrastive study of cognate noun/verb pairs in English and Norwegian

Jarle Ebeling and Signe Oksefjell Ebeling
(University of Oslo)

Previous cross-linguistic studies of specific cognate word pairs in English and Norwegian have shown varying degrees of overlap in frequency and use across the languages, e.g. Ebeling (2017) on *bring* and *bringe* and Ebeling (2024) on *see* and *se*. This paper widens the scope and investigates a set of cognate English and Norwegian heterosemeous words that, through conversion, "express related meanings across multiple word classes" (Shao et al. 2023: 321), e.g. *star* (N) and *star* (V). In a diachronic study of heterosemy in recent English (1920s-2010s), Shao et al. (2023) compiled a list of 877 heterosemous nouns and verbs from the Corpus of Historical American English. This list will serve as the starting point of the present English-Norwegian contrastive analysis, with a focus on pairs that have etymologically related counterparts in Norwegian.

In contrast to English, conversion in Norwegian often requires some affixation, e.g. *hat* (N) vs. *hate* (V) 'hate', although zero-derivational pairs also exist, e.g. *pumpe* (N) and *pumpe* (V) 'pump'. Thus, conversion by zero-derivation was not a requirement in the Norwegian material. This is in line with Huddleston & Pullum's (2002) understanding of conversion.

After identifying corresponding cognate noun/verb pairs, we aim to establish how and to what extent such pairs are used in English and Norwegian. We do this by searching for each pair in the English-Norwegian Parallel Corpus+, a bidirectional corpus that will enable us to establish the items' Mutual Correspondence (Altenberg 1999). In this investigation of more than 100 noun and verb pairs we will be in a position to provide a more comprehensive and accurate account of how cognates behave cross-linguistically. To our knowledge, a study of cognate heterosemeous noun/verb pairs has not been done on this scale before, as most previous studies have focused on individual cognate pairs, often belonging to one word class only.

More specifically, we seek answers to the following research questions:

1. To what extent do noun and verb uses of the cognates correspond to each other in translation?
2. What factors tend to trigger a higher or lower degree of Mutual Correspondence (MC)?

Preliminary results show that the nouns generally tend to have a higher MC than the verbs, indicating that the verbs more commonly develop diverging polysemies and/or conditions of use. It is expected that these findings will be substantiated in the analysis of the complete set of noun/verb pairs.

Although similar trends are reported at the overall level, cognate noun/verb pairs show contrastive differences at the individual level in terms of frequency and MC. The final part of the study, addressing the second research question, will therefore consist of case studies analysing individual pairs that show different degrees of MC in the hope of revealing features that may contribute to this variation in MC between two closely related languages. While MC is meant as an initial step in establishing cross-linguistic equivalence, a more detailed cross-linguistic analysis of the kind proposed here is necessary to shed further light on precisely what is common to the languages compared and what sets them apart.

References

Altenberg, Bengt (1999), Adverbial connectors in English and Swedish: Semantic and lexical correspondences, in Hilde Hasselgård and Signe Oksefjell (eds), *Out of Corpora: Studies in Honour of Stig Johansson*, Amsterdam: Rodopi, 249–268.

Ebeling, Signe Oksefjell (2017), Bringing home the bacon! A contrastive study of the cognates *bring/bringe* in English and Norwegian, *Kalbotyra* 70, 104–126.

Ebeling, Signe Oksefjell (2024), Seeing through Languages and Registers: A Closer Look at the Cognates *See* and *Se*, in Anna Cermakova, Hilde Hasselgård, Markéta Malá and Denisa Šebestová (eds), *Contrastive Corpus Linguistics. Patterns in Lexicogrammar and Discourse,* London: Bloomsbury Academic, 29–61.

Huddleston, Rodney and Geoffrey K. Pullum (2002), *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press.

Shao, Bin, Jing Zheng, and Hendrik De Smet (2023), The blurring of the boundaries: changes in verb/noun heterosemy in Recent English, *Corpus Linguistics and Linguistic Theory,* 20(2), 321–346.

Verb insertion in translations as a sign of grammaticalisation in progress

Thomas Egan
(University of Inland Norway)

When analysing translations of expressions in parallel corpora, one normally distinguishes between syntactically congruent and non-congruent translations. There may also be expressions in the original text which the translator has omitted altogether, or expressions in a translation lacking a correspondence in the original. These are both referred to as 'zero translations' (see Johansson 2007: 26). It is the second type of zero translation that is the topic of this presentation. An example of such a translation is (1), taken from the English–Norwegian Parallel Corpus (ENPC), in which the the ingressive aspect signalled by *begynne* (begin) in the translation is not present in the English original.

(1) Inside his head the synapses were shutting down. (DF1)
I hodet *begynte* synapsene å slukne. (DF1T)
'In the head began the synapses to go out.'

When an expression is undergoing grammaticalisation, it normally undergoes semantic bleaching (see, for example, Heine and Narrog (2010: 406)). A textbook example of a grammaticalised expression in English is the *going to* future, as in (2).

(2) "Jeg *skal* møte henne imorgen." (LSC2)
'"I shall meet her tomorrow."'
"I'm *going to* see her tomorrow." (LSC2T)

In (2) the expression *going to* does not translate a verb of inherently directed motion (Levin 1993: 263), but the present tense form of the Norwegian modal verb *skulle*, here coding a planned action (Faarlund et al. 1997: 604). The translation is syntactically congruent, since in both the source and target text the first verb licences an infinitive clause.

(1) and (2) are, respectively, clear-cut examples of a zero translation and a congruent translation of a grammaticalised expression. In this presentation I examine less clear-cut cases, investigating a handful of constructions which may be in the process of grammaticalising. These include English constructions containing the matrix verbs *help* (Mair 1995) and *fail* (Mackenzie 2008, Egan 2016) and Norwegian pseudo-coordinate constructions headed by the posture verbs *stå* (stand) and *sitte* (sit) (Kinn *et al*. 2018, among many others). All of these constructions have already been subject to some contrastive analysis, the *fail* construction in Egan (2018), the *help* construction in Egan (2024), and the Norwegian constructions in

Tonne (1999), Johansson (2009) and Ebeling (2015). Here I revisit the constructions, on the basis of data from one and the same corpus, the ENPC, and consider all examples in the translations in which the first verb, be it matrix verb or first coordinate, lacks a lexical correspondence in the original text. The research questions is as follows:

Can the insertion of a second verb in the translation of a one-verb construction be taken as an indication of the grammaticalisation of (that verb in) the resulting two-verb construction?

Preliminary analysis shows that the corpus contains enough evidence on which to base the discussion of this question.

References

Ebeling, Signe Oksefjell (2015), A contrastive study of Norwegian pseudo-coordination and two English posture-verb constructions, in Signe Oksefjell Ebeling and Hilde Hasselgård (eds), *Cross-Linguistic Perspectives on Verb Constructions*, Newcastle: Cambridge Scholars Publishing, 29–57.

Egan, Thomas (2016), The subjective and intersubjective uses of FAIL TO and 'not fail to', in Hubert Cuyckens, Lobke Ghesquière and Daniel van Olmen (eds), *Aspects of Grammaticalization: (Inter)subjectification and Pathways of Change*, Berlin: De Gruyter, 168–196.

Egan, Thomas (2018), The FAIL TO Construction: A Contrastive Perspective, *Bergen Language and Linguistics Studies,* 9(1). https://doi.org/10.15845/bells.v9i1.1525.

Egan, Thomas (2024), Simple and complex help constructions in English and Norwegian: a contrastive study, *Languages in Contrast,* 24(1), 84–108.

Faarlund, Jan Terje, Svein Lie and Kjell Ivar Vannebo (1997), *Norsk Referansegrammatikk*, Oslo: Universitetsforlaget.

Heine, Bernd and Heiko Narrog (2010), Gerammaticalization and Linguistic Analysis, in Bernd Heine and Heiko Narrog (eds), *The Oxford Handbook of Linguistic Analysis,* Oxford: Oxford University Press, 401–423.

Johansson, Stig (2007), *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*, Amsterdam: John Benjamins.

Johansson, Stig (2009), Norwegian-pseudo-coordination with verbs of posture in translation into English and German, in Stefan Slembrouck, Miriam Taverniers and Mieke Van Herreweghe (eds), *From* Will *to* Well: *Studies in Linguistics offered to Anne-Marie Simon-Vandenbergen*, Ghent: Academia Press, 279–291.

Kinn, Torodd, Kristian Blensenius, and Peter Andersson (2018), Posture, location, and activity in Mainland Scandinavian pseudocoordinations, *Cognitextes* 18. *https://doi.org/10.4000/cognitextes.1158*

Levin, Beth (1993), *English verb Classes and Alternations: A Preliminary Investigation,* Chicago: University of Chicago Press.

Mackenzie, J. Lachlan (2008), Failing without trying, *Jezikoslovlje* 9(1-2), 53–85.

Mair, Christian (1995), Changing patterns of complementation, and concomitant grammaticalisation, of the verb help in present-day English, in Bas Aarts and Charles F. Meyer (eds), *The Verb in Contemporary English: Theory and Description*, Cambridge: Cambridge University Press, 258–72.

Tonne, Ingebjørg (1999), A Norwegian progressive Marker and the Level of Grammaticalization, *Languages in Contrast* 2(1), 131–159.

Verbal expressions of habituality in English and Norwegian: Forms and correspondences

Hilde Hasselgård
(University of Oslo)

Habitual expressions "describe a situation which is characteristic of an extended period of time, so extended in fact that the situation referred to is viewed not as an incidental property of the moment but, precisely, as a characteristic feature of a whole period" (Comrie 1976: 27-28). While there is no fully

grammaticalized habitual aspect in verb systems of English and Norwegian, both languages have periphrastic expressions of habitual meaning as defined above. Examples are *used to, tend to* (*I used to help him out / This tends to work well*) and *pleie, bruke* (*Faren pleide å gi ham en bok… – His father used to give him a book*…). Besides catenative expressions, modal auxiliaries (e.g. English *would,* Norwegian *kunne*) and complement clause constructions can express habituality, e.g. Norwegian *det hender at* 'it happens that' and English *be known to*. Previous comparisons of English and Norwegian have focused on a single expression of habituality and its correspondences, e.g. Bjerga (1998), who studied Norwegian *pleie* and its English correspondences, and Johansson's (2005) examination of *det hender at* 'it happens that'. Lund (2007) investigated whether *used to* could always be translated by *pleie*. These studies found that habitual expressions are rarely translated congruently between English and Norwegian. Similarly, Altenberg (2007) found low mutual correspondence between English *used to* and Swedish *bruka.*

The present study considers verbal expressions of habituality in both Norwegian and English and their correspondences. The material comes from the English-Norwegian Parallel Corpus, in which the following expressions have been searched for in original texts: the habitual catenatives *pleie* and *bruke* + infinitive and the English expressions *used to, tend to, be known to* and *would*. The aims are the following: to compare the use and distribution of habituality expressions within and between the languages; to survey the range of translation correspondences of each expression; and to identify potential factors influencing the choice of correspondence.

Preliminary results show that *pleie* far outnumbers *bruke* in Norwegian originals. In English *would* is most frequent, followed by *used to, tend to* and *be known to*. In sum, the English habituals are more frequent than the Norwegian ones. Recurrent noncongruent correspondences include adverbials, e.g. *usually, sometimes, alltid* ('always') *før* ('before'). This aligns with Altenberg's (2007) results, and was also apparent in Hasselgård (2007), in which usuality adverbials sometimes corresponded to verbal expressions. Furthermore, translations into both languages frequently omit the habituality marker and use a simple tense instead, and Norwegian may use the perfect aspect to mark anteriority. Altenberg (2007) shows that co-occurring time adverbials and the tense and dynamicity of the verb phrase influence the choice of translation between English and Swedish. These factors will also be examined in the proposed study, which in addition aspires to build on and connect the findings of the previous studies and to identify a possible division of labour between verbal, adverbial and other expressions of habituality in both languages examined.

References

Altenberg, Bengt (2007), Expressing past habit in English and Swedish, in Christopher Butler, Raquel Hidalgo, Julia Lavid Downing (eds), *Functional Perspectives on Grammar and Discourse: In Honour of Angela Downing*, Amsterdam: John Benjamins, 97–128.

Bjerga, Trude Davidsen (1998), Continuative and habitual aspect in English and Norwegian. With special reference to the English verb *keep* and the Norwegian verb *pleie*, Unpublished MA thesis, University of Oslo.

Comrie, Bernard (1976), *Aspect*, Cambridge: Cambridge University Press.

Hasselgård, Hilde (2007), Using the ENPC and the ESPC as a parallel translation corpus: Adverbs of frequency and usuality, *Nordic Journal of English Studies*, 6(S1). https://doi.org/10.35360/njes.8

Johansson, Stig (2005), Some aspects of usuality in English and Norwegian, in Eva Maagerø and Kjell Lars Berge (eds), *Semiotics from the North: Nordic Approaches to Systemic Functional Linguistics*, Oslo: Novus, 69–86.

Lund, Karianne (2007), Kan 'used to' alltid oversettes med 'pleide'?: en sammenlikning av used to og pleie som uttrykk for habitualitet, med fokus på oversettelse fra engelsk til norsk, MA thesis, University of Oslo. http://urn.nb.no/URN:NBN:no-16964.

Corpus

The English-Norwegian Parallel Corpus, see https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/enpc/

Shifts from noun phrase postmodification to premodification in academic writing: Towards conditions and contexts of change

Marie-Pauline Krielke and Isabell Landwehr
(Saarland University)

Relative clauses (RCs) provide explicit noun phrase (NP) post-modification and over time, they have become less favored in scientific English. Facilitated by greater reliance on background knowledge, they have been replaced by denser, less explicit NP modifications like attributive adjectives and compounds (Hundt *et al.*, 2012; Biber & Gray, 2016),. In German, RCs initially increase before declining later than in English, reflecting German's delayed establishment as a primary scientific language (Krielke, 2021). However, RCs persist in both languages. This study investigates when RCs are replaced and when they survive in scientific discourse, focusing on RCs with copula verbs and predicative adjectives (e.g., "*the element which is solid*," RC+A). Their development is compared to attributive adjective + noun constructions (e.g., "*the solid element*," A+N). An information-theoretic approach guides this analysis, based on the following assumptions:

a. A+N usage increases in both English and German, while RC+A declines in English but initially rises and later declines in German.

b. A+N usage becomes increasingly predictable (lower surprisal) compared to RC+A in both languages.

c. Surviving RC+A constructions become more predictable over time, occurring in entrenched (Bybee, 2002) contexts where grammatical necessity prevents more condensed expressions.

Our data set for English is the Royal Society Corpus, (RSC Version 6.0 Open, Fischer et al., 2020) consisting of texts from the Philosophical Transactions and Proceedings of the Royal Society between1650 – 1920. The German data set is the scientific portion of Deutsches Textarchiv (DTAW, Geyken *et al.,* 2018) comprising scientific books covering 1600 – 1890. Both datasets are annotated with parts-of-speech and 4-gram surprisal (Shannon, 1948). The information-theoretic notion of surprisal captures a word's predictability given its context (here: the three preceding words). High surprisal indicates low predictability and vice versa.

First results show that A+N constructions increase steeply in both languages, while RC+A constructions are less frequent. In English, RC+A constructions decline, whereas in German, they initially rise and then decrease after 1740.

Surprisal of RC+A adjectives is constantly higher compared to A+N in both languages, likely due to their lower frequency. Despite declining frequency in English, RC+A surprisal decreases, indicating increasingly entrenched usage: Surviving RC+A constructions feature entrenched forms with further specifications (e.g., prepositional phrases or infinitival complements), while simple "*x, which is y*" constructions become rare. In German, surprisal increases for both constructions, suggesting less entrenched adjective usage. Although adjectives directly followed by "*sein*" decrease, the variability (entropy, Shannon, 1948) of other continuations remains stable. These findings support the hypothesis that, in expert scientific communication, such simple constructions are avoided to reduce redundancy. Their partial survival is motivated by their inability to be prepended, unlike standalone adjectives.

Future analyses will explore whether RC+A constructions in highly specialized contexts (e.g., "*the plane which is perpendicular"*) serve as precursors to A+N constructions and examine the role of surprisal in driving their shift toward denser encodings. Specifically, we will assess whether decreasing surprisal in RC+A constructions facilitates the emergence of corresponding A+N forms.

References

Biber, Douglas and Bethany Gray (2016), *Grammatical complexity in academic English: Linguistic change in writing*, Cambridge University Press.

Bybee, Joan L. (2002), Main clauses are innovative, subordinate clauses are conservative: consequences for the nature of constructions, in Joan Bybee and Michael Noonan (eds), *Complex sentences in grammar*

and discourse: essays in honor of Sandra A. Thompson, Amsterdam/Philadelphia: John Benjamins, 1–18.

Fischer, Stefan, Jörg Knappen, Katrin Menzel and Elke Teich (2020), The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study, in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, 794-802.

Geyken, Alexander, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, and Frank Wiegand (2018), Das Deutsche Textarchiv als Forschungsplattform historische Daten in CLARIN, in Henning Lobin, Roman Schneider and Andreas Witt (eds), *Digitale Infrastrukturen die germanistische Forschung* (=Germanistische Sprachwissenschaft um 2020, Bd. 6), Berlin/Boston: De Gruyter, 219–248.

Hundt, Marianne, David Denison, and Gerold Schneider (2012), Relative complexity in scientific discourse, *English Language and Linguistics*, 16(2), 209–240.

Krielke, Marie-Pauline (2021), Relativizers as markers of grammatical complexity: A diachronic, cross-register study of English and German, *Bergen Language and Linguistics Studies*, 11(1), 91–120.

Shannon, Claude Elwood (1948), A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. The Bell System Technical Journal.

Swedish compound nouns and English noun sequences – a perfect match?

Magnus Levin, Jenny Ström Herold and Vasiliki Simaki
(Linnaeus University, Lund University)

In a previous cross-linguistic study (Ström Herold & Levin, forthcoming), we investigated English noun sequences, i.e. juxtaposed nouns (*world war*; *health care reform initiative*), with their German and Swedish correspondences, the results indicating that around 70% of the correspondences are (solid) noun-noun compounds (*Weltkrieg/världskrig*), irrespective of language and translation direction. Thus, there appears to be a fairly strong cross-linguistic correlation between these two types of structures – at least with this methodological approach. What we do not know at present is if this observed correlation is an interference effect from English. To address this knowledge gap, we reverse the study, instead taking Swedish compounds as the point of departure.

Therefore, this study aims to explore: (i) the proportion of Swedish compound nouns in relation to their English correspondences (noun sequences or compounds), (ii) the distributions of other types of correspondences such as postmodifying prepositional phrases and premodifying adjectives, and (iii) what these results tell us about language-specific preferences and translation effects.

In this study, an extended tagged version of the Linnaeus University English-German-Swedish corpus (LEGS) non-fiction corpus is used (Ström Herold & Levin 2019; forthcoming). LEGS consists of, e.g., popular science, biographies and self-help books amounting to approximately half a million words of each source language. Using a Python script, we retrieve words tagged as nouns (and their translations), which are then classified, manually removing all non-compound nouns.

We will be considering the same variables as in our previous study: compound noun length (two-part, three-part etc.), common vs. proper nouns as first elements (e.g., *law degree/juristexamen* vs. *Yale degree/Yaleexamen*) and the semantic relations holding between the parts of the nouns (e.g., time relation, purpose relation; cf. Teleman et al. 1999: II: 44–45 for Swedish, and Biber et al. (2021 [1999]: 582) for English). As for the first variable, a corpus study by Carlsson (2004: 75), contrasting Swedish and German newspaper language, showed that two-part compounds constitute more than 90% of her material.

Our preliminary findings suggest that Swedish compounds are rendered as English noun sequences (*familjemedlemmar* ['family-members'] > *family members*) or solid compounds (*grundvatten* ['ground-water'] > *groundwater*) in proportions similar to those of the previous study. As for 'non-compound' correspondences, English appears to use slightly more premodifying adjectives as correspondences to Swedish compounds (*flingsalt* ['flake-salt'] > *flaked salt*) than in our previous study. Swedish compound

nouns seem to be shorter than English noun sequences and less frequently contain proper nouns as first elements, as compared to findings in Ström Herold & Levin (2019; forthcoming). Finally, 'simple' English noun correspondences (*vetemjöl* ['wheat-flour'] > *flour*) are rare (cf. the high frequencies identified by Nesset (2018) in Norwegian-Russian contrast).

Our study will deepen the state of knowledge regarding the similarities and differences between Swedish and English noun phrase structures, an area that is still under-researched for this language pair. It will also try to disentangle the effects of source-language norms or restrictions and translation-induced changes.

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (2021 [1999]). *Longman Grammar of Spoken and Written English*, Harlow: Longman.

Carlsson, Maria (2004), *Deutsch und Schwedisch im Kontrast: Zur Distribution nominaler und verbaler Ausdrucksweise in Zeitungstexten* [Göteborger germanistische Forschungen 43], Göteborg: Acta Universitatis Gothoburgensis.

Nesset, Tore (2018), When a single word is enough: Norwegian compounds and their Russian counterparts, *Slovo. Journal of Slavic Languages, Literatures and Cultures,* 59, 61–72.

Ström Herold, Jenny and Levin, Magnus (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 Herold, Jenny and Levin, Magnus (Forthcoming), *Climate change* and *Harvard students*: English noun sequences and their German and Swedish correspondences, to appear in *Corpora,* 18(3).

Teleman, Ulf, Erik Andersson and Staffan Hellberg (1999), *Svenska Akademiens Grammatik*, Volume II, Stockholm: Norstedts.

Complexity of the noun phrase in English and Czech children's literature

Markéta Malá
(Charles University)

While children's literature is often considered a simplified version of fiction for adults (Thompson & Sealey 2007), there appear to be relatively few studies actually measuring grammatical complexity of fiction for children (e.g., Puurtinen 1998, Montag 2019, Dawson 2023). Complexity measures have been used to analyse genre and register differences (Biber et al. 2011, 2024), the development of children's writing and reading proficiency (Hsiao et al. 2023, 2024), and especially to assess L2 students' proficiency in English (e.g. Bulté & Housen 2012).

This work-in-progress study sets out to explore some of the factors which contribute to syntactic complexity in children's literature in two typologically distinct languages, English and Czech. I focus on the structure of noun phrases headed by nouns, which appears to mirror the developmental stages in language proficiency, with heavy postmodification (especially by prepositional phrases and nonfinite clauses) being indicative of high levels of complexity (Biber et al. 2011).

The research draws on data from the English and Czech sections of the parallel corpus InterCorp. Its latest edition has been annotated using the universal dependencies framework (Rosen 2023) and several measures of syntactic complexity (at the level of the sentence and text) and lexical diversity. The replacement of language-specific morphological tagging by language-uniform morpho-syntactic annotation now makes it possible to expand the scope of contrastive studies to include direct comparison of syntactic structure and complexity across languages. The study is based on two small comparable sub-corpora (0.5 million tokens each) of InterCorp comprising English and Czech original fiction for children.

In both languages, noun phrases in children's fiction were found to be generally shorter, with fewer layers of embedding than in fiction for adults. At the same time, noun phrases in English children's fiction

are, on average, longer than those in Czech children's books (3.2 and 3.35 words per phrase, respectively), with similar maximum depth of embedding. This suggests the impact of the analytical character of English as opposed to predominantly synthetic Czech: the category of definiteness typically remains unexpressed in Czech, and the relations between the head and a postmodifying noun may be expressed by case suffixes rather than by prepositions.

As shown by Biber et al. (2024), however, apart from structural distinctions, syntactic functional distinctions have to be considered when studying complexity. The preliminary results indicate that a higher proportion of noun-headed phrases function as the subject in Czech than in English. The fact that the subject can occupy the clause-final position in Czech may then contribute to heavier postmodification within the Czech subject noun phrases. The two languages were also found to differ in the forms of postmodification. The higher proportion of non-finite clauses in English, as opposed to finite relative clauses, can be interpreted in relation to the preference of Czech for finite verb predicates, which make it possible to express the verbal grammatical categories on the lexical verb (cf. Dušková 2015).

References

Biber, Douglas, Bethany Gray and Kornwipa Poonpon (2011), Should We Use Characteristics of Conversation to Measure Grammatical Complexity In L2 Writing Development?, *TESOL Quarterly,* 45(1), 5–35.

Biber, Douglas, Tove Larsson and Gregory R. Hancock (2024), The linguistic organization of grammatical text complexity: comparing the empirical adequacy of theory-based models, *Corpus Linguistics and Linguistic Theory,* 20(2), 347–373.

Bulté, Bram and Alex Housen (2012), Defining and Operationalising L2 Complexity, in Alex Housen, Folkert Kuiken and Ineke Vedder (eds), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*, Amsterdam/Philadelphia: John Benjamins.

Corpus InterCorp, version 16ud from 17 September 2024. Ústav Českého národního korpusu FF UK, Praha. Available from http://www.korpus.cz

Dawson, Nicola, Ya-Ling Hsiao, Alvin Wei Ming Tan, Nilanjana Banerji and Kate Nation (2023), Effects of Target Age and Genre on Morphological Complexity in Children's Reading Material, *Scientific Studies of Reading*, 27(6), 529–556.

Dušková, Libuše (2015), *From Syntax to Text: the Janus Face of Functional Sentence Perspective*, Praha: Karolinum.

Hsiao, Yaling, Nicola J. Dawson, Nilanjana Banerji, and Kate Nation (2023), The nature and frequency of relative clauses in the language children hear and the language children read: A developmental cross-corpus analysis of English complex grammar, *Journal of Child Language,* 50(3), 555–580.

Hsiao, Yaling, Nicola J. Dawson, Nilanjana Banerji, and Kate Nation (2024), A corpus-based developmental investigation of linguistic complexity in children's writing, *Applied Corpus Linguistics*, 4(1), 1–14.

Montag, Jessica L. (2019), Differences in sentence complexity in the text of children's picture books and child-directed speech, *First Language*, 39(5), 527–546.

Puurtinen, Tiina (1998), Syntax, Readability and Ideology in Children's Literature, *Meta* 43(4), 524–533.

Rosen, Alexandr (2023), The InterCorp parallel corpus with a uniform annotation for all languages, *Jazykovedný časopis*, 74(1), 254–265.

Thompson, Paul and Alison Sealey (2007), Through Children's Eyes? Corpus Evidence of the Features of Children's Literature, *International Journal of Corpus Linguistics,* 12(1), 1–23.

Academic voices past and present: First-person pronouns in English and Norwegian research articles

Sylvi Rørvik
(University of Inland Norway)

This study examines the frequency and rhetorical functions of first-person pronouns in English and Norwegian research articles (RAs). Numerous studies have investigated the use of first-person pronouns as indications of authorial presence in academic texts in a range of languages, including English, Spanish, Swedish, and Lithuanian, as well as many others (cf. e.g. Hyland 2001, Sheldon 2009, McGrath 2016, Hyland & Jiang 2017, Šinkūnienė 2018, Carrió-Pastor 2020, Wheeler et al 2021, Ädel 2022, and Dixon 2022). However, Norwegian academic prose has not been extensively studied. One seminal exception is Fløttum et al's work from 2006, which examined the use of various features including first-person pronouns in Norwegian, French, and English RAs (dating from the early 1990s to the early 2000s) from three fields: economics, linguistics, and medicine. Fløttum et al identified disciplinary differences in the use of first-person pronouns.

Using Fløttum et al's study as the starting point, the present study investigates English and Norwegian RAs from the fields of linguistics and education, including 50 RAs in each field in each language, i.e. 200 in total. These articles have all been peer-reviewed and accepted for publication, and are therefore considered to be representative of English and Norwegian academic prose, even if the native-speaker status of the authors has not been ascertained (cf. Carrió-Pastor 2020: 19).

The aim of the study is partially to enable a diachronic comparison with Fløttum et al's results by employing more recent linguistics material (dating from 2015-2024), but the inclusion of RAs from the field of education serves a primarily pedagogical aim: to provide an empirical basis for teaching and supervising students in the field of education who are required to write academic papers in both English and Norwegian, which is the case for Norwegian students who aim to become teachers of English. The research questions are as follows:

1. To what extent are there cross-linguistic and cross-disciplinary differences in the *frequency* of first-person subject pronouns in English and Norwegian RAs in the fields of linguistics and education?
2. To what extent are there cross-linguistic and cross-disciplinary differences in the *rhetorical functions* of first-person subject pronouns in English and Norwegian RAs in the fields of linguistics and education?

Preliminary results regarding the frequency of first-person pronouns indicate that the subcorpora are characterized by within-group variation, but that both language and discipline seem to play a role in the use of first-person pronouns: In the English material, first-person plural subject pronouns are markedly more frequent in linguistics than in education. In Norwegian the disciplines are very similar, but the frequency for education is twice as high as that observed for English, and for linguistics it is approximately 1/3 higher than that found in the English material. The presentation will also include an overview of the rhetorical functions expressed by the first-person subject pronouns in the material. This analysis builds on the four roles outlined by Fløttum et al (2006: 81, 83-84): writer, researcher, arguer, and evaluation.

References

Ädel, Annelie (2022), Writer and reader visibility in humanities research articles: Variation across language, regional variety and discipline, *English for Specific Purposes,* 65, 49–62.

Carrió-Pastor, María Luisa (2020), Variations in the Use of Self-Mentions in Different Specific Fields of Knowledge in Academic English, in María Luisa Carrió-Pastor (ed), *Corpus Analysis in Different Genres: Academic Discourse and Learner Corpora,* Taylor & Francis Group, 13–32.

Dixon, Tulay (2022), Proscribed informality features in published research: A corpus analysis, *English for Specific Purposes,* 65, 63–78.

Fløttum, Kjersti, Trine Dahl and Torodd Kinn (2006), *Academic Voices: Across languages and disciplines,* John Benjamins Publishing Company.

Hyland, Ken (2011), Humble servants of the discipline? Self-mention in research articles, *English for Specific Purposes,* 20(3), 207–226.

Hyland, Ken and Feng Jiang (2017), Is academic writing becoming more informal? *English for Specific Purposes,* 45, 40–51.

McGrath, Lisa (2016), Self-mentions in anthropology and history research articles: Variation between and within disciplines, *Journal of English for Academic Purposes,* 21, 86–98.

Sheldon, Elena (2009), From one *I* to another: Discursive construction of self-representation in English and Castilian Spanish research articles, *English for Specific Purposes,* 28(4), 251–265.

Šinkūnienė, Jolanta (2018), *I* and *we* in Lithuanian, Lithuanian English and British English research writing, in Pilar Mur-Dueñas (ed), *Intercultural Perspectives on Research Writing,* John Benjamins Publishing Company, 59–79.

Wheeler, Melissa A., Ekaterina Vylomova, Melanie J. McGrath and Nick Haslam (2021), More confident, less formal: stylistic changes in academic psychology writing from 1970 to 2016, *Scientometrics,* 126, 9603–9612.

Contrastive analysis of English and German in a construction grammar framework

Peter Uhrig and Thomas Herbst
(University of Erlangen-Nuremberg)

We attempt to show how a Construction Grammar (CxG) framework (Goldberg 2019; Hoffmann 2022; Herbst & Hoffmann 2024) can be used to carry out analyses of two closely related languages – German and English. We will build on earlier work on contrastive linguistics (e.g. Burgschmidt & Götz 1974; König & Gast 2018) in that we regard it as essential that any contrastive study should be based on descriptions of the two languages that were carried out independently from one another so as not to prioritize one language over the other. In our view, the model of CxG is particularly well suited for contrastive analyses because it is based on constructions as form-meaning pairings, although we would argue that in the field of contrastive analysis it would be wrong to say that the semantic side of grammatical constructions had been neglected in the structuralist work of the 1960s and 1970s. In contrast to at least some of the very early approaches within contrastive linguistics, we would not argue that contrastive analyses would enable one to predict errors L2-learners are likely to make.

The area of description we intend to focus on is that of argument structure constructions in English and German, which was captured in a lot of descriptive work in the form of complementation patterns (Quirk et al. 1985, and also Hunston & Francis 2000) for English and in terms of valency for German (e.g. Helbig & Schenkel 1968 or Schumacher et al 2004). We will address issues such as the following:

To what extent can one identify corresponding, equivalent or parallel constructions in the two languages, if one assumes an approach that sees constructions as language specific? We would argue that we can identify a double object (i.e. ditransitive) construction in German and in English and that there are clear parallels between them in that they have 4 slots: NP (subject) – V – NP (indirect object) – NP (direct object) with comparable roles (agent or æffector – action – recipient – theme/Patient/æffected), which allow us to establish a certain degree of correspondence between the German and English ditransitive constructions. However, the identification of the three different NP slots is driven by word order in English and by case in German. We are going to critically discuss Croft's (2022) comparative concepts in the light of this comparison. Another difference between the German and the English ditransitive constructions is revealed by comparing their collo-profiles, i.e. a corpus-based and frequency-related analysis of the verbs that occur in them (cf. www.constructicon.de; Herbst & Hoffmann 2024). On the basis of different English and German argument structure constructions, we will argue that the comparison of such collo-profiles is a key element of a contrastive analysis in this area, which, of course, was impossible to carry out in pre-corpus times, and which reveals important information about the difference between two languages that – combined with empirical error analysis making use of learner corpora, for instance – can be exploited in the design of foreign language teaching materials and textbooks.

References

Burgschmidt, Ernst and Dieter Götz (1974), *Kontrastive Linguistik Deutsch/Englisch*, München: Hueber.

CASA|Con (2023-), *A Constructicon of the English Language*, Thomas Herbst, Thomas Hoffmann, Peter Uhrig, Armine Garibyan and Stephanie Evert (eds). www. constructicon.de.

Croft, William (2022), *Morphosyntax: Constructions of the World's Languages*, Cambridge: Cambridge University Press.

Goldberg, Adele E. (2019), *Explain Me This,* Princeton: Princeton University Press.

Helbig, Gerhard and Wolfgang Schenkel (1968), *Wörterbuch zur Valenz und Distribution deutscher Verben,* Leipzig: Enzyklopädie.

Herbst, Thomas and Thomas Hoffmann (2024), *A Construction Grammar of the English Language. CASA – a constructionist approach to syntactic analysis,* Amsterdam/Philadelphia: Benjamins.

Hoffmann, Thomas (2022), *Construction Grammar: The Structure of English,* Cambridge: Cambridge University Press.

Hunston, Susan and Gill Francis (2000), *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*, Amsterdam/Philadelphia: Benjamins.

König, Ekkehard and Volker Gast (2018), *Understanding English-German Contrasts*, 4th edition, Berlin: Erich Schmidt Verlag.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1985), *A Comprehensive Grammar of the English Language*, London: Longman.

Schumacher, Helmut, Jacqueline Kubczak, Renate Schmidt and Vera de Ruiter (2004), *VALBU – Valenzwörterbuch deutscher Verben*, Tübingen: Narr.

General Sessions

Searching for Old English influence in Old Norse homilies

Karoline Aastrup-Köhler
(University of Oslo)

Work-In-Progress

While much work has been done on the influence of Old Norse on English during the Viking Age, relatively little has been written on linguistic influence in the other direction. A noteworthy exception is research on lexical borrowing of religious terminology due to the Christianisation of Norway and the organisation of the Norwegian church, both processes predominantly led from England (Gunn 2017). And this is the starting point of my PhD project: If English clergymen influenced the vocabulary of the newly founded Norwegian church, it is not farfetched to also assume that Old English influenced the Old Norse language in other ways, such as through pragmatic borrowing (Prince 1988), and that this influence would be particularly visible in Old Norse learned/religious texts that have been shown to have an Anglo-Saxon provenance – such as the sermons in the *Old Norwegian Homily Book* (Abram 2004). My project investigates the use of verb-initial word order (V1), both in Old English and in Old Norse – including the linguistic factors that led to its use, as well as its pragmatic functions. This word order pattern was chosen because there exists sufficient relevant research to build on for both languages. My hypothesis is that in the Old Norse homilies, the use of V1 will be more similar to the Old English use than its use in e.g. saga texts is – particularly as concerns pragmatic functions, possibly also as concerns other linguistic factors.

Research questions:

[1] What linguistic factors facilitate the use of marked V1 word order in Old English vs. in Old Norse?

[2] To what extent can we find Old English influence on the use of V1 word order in the *Old Norwegian Homily Book*?

To answer question [1], I will conduct a corpus-based, quantitative study of V1 in Old English and Old Norse prose. I will use three corpora: the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (YCOE), the *Menotec* corpus, and the *Icelandic Parsed Historical Corpus* (IcePaHC). To answer question [2], I will then perform a qualitative, in-context analysis of a smaller number of V1 and non-V1 sentences from two Old Norse works – the *Old Norwegian Homily Book* and a saga text.

As for the expected results: Whether or not my hypothesis is confirmed, my work will result in a comparative examination of the use of V1 word order in Old English and Old Norse prose, and possibly new insight into English–Norwegian ecclesiastical contact in the late Viking Age and early Scandinavian Middle Ages. Moreover, my hope is that this project will serve as a starting point for future research into Old English influence on Old Norse.

By June 2025, I expect to be well underway with the quantitative study and the mapping out of the effects of linguistic factors such as verb type, subject type, and the presence of different coordinating conjunctions on the use of marked V1; these results will be central to my work-in-progress report.

References

Abram, Christopher (2004), 'Anglo-Saxon Influence in the Old Norwegian Homily Book', *Mediaeval Scandinavia* 14(a), 1–35.

Gunn, Nikolas (2017), *Contact and Christianisation: Reassessing Purported English Loanwords in Old Norse* (Doctoral dissertation), University of York.

Prince, Ellen F. (1988), 'On Pragmatic Change: The Borrowing of Discourse Functions', *Journal of Pragmatics* 12(5-6), 505–518.

'…no one knows what's on the other side': Oppositionality in the discourse of Scottish opinions on assisted dying

Marc Alexander and James Balfour
(University of Glasgow)

The current Assisted Dying for Terminally Ill Adults (Scotland) Bill aims to enable mentally competent, terminally ill adults to be provided, at their request, with medical assistance to end their life. The issue in Scotland has remained contentious, with faith groups generally opposing change while polls consistently showing public support above 70%. Given the serious nature of the topic, opinions on this subject are frequently personal and deeply passionate. The public were consulted on the bill twice, in 2022 and 2024, and these responses are valuable sources of authentic short texts from members of the public on a highly emotive topic.

This presentation will focus on the polarisation of the discourse surrounding this debate, contrasting the responses from those who self-identify as either supporting or opposing the bill. We aim to contribute not only to the understanding of the discourse of this particular debate, but also to the more general question of how best to analyse oppositional discourse in public life.

We built corpora consisting of 2.1m words from 12,314 written responses to a consultation from the proposing Member of the Scottish Parliament, closing in 2022, and 1.8m words from 7,236 written responses to a consultation by the Parliament's Health, Social Care and Sport Committee in late 2024. These were semantically tagged using WMatrix, and the metadata for the corpus contained for each response the degree of support or opposition the author reported for the bill.

To analyse the rhetoric of positioning and dissent on this topic, we will present work on four areas. First, we look at the 'unique' lexis – that is, lexis not shared between our supportive and opposing subcorpora – such as *hideous, abject, urine, linger* for supportive writers against *eroded, burden, wedge, shalt* for those opposed. We then analyse the frequency within each subcorpus of items which do not overlap with opposing viewpoints. Secondly, we compare the ranking of key semantic domains for our subcorpora (log-likelihood against a standard reference corpus) against each other, analysing which domains are substantially more or less commonly used in each subcorpus, such as the high ranking of *E4.1-(Sad)* for supportive responses compared to opposed, while the opposite pattern appears for *A7+ (Likely)*, showing the higher emotional content of supportive authors compared to the more analytic responses of many of those opposed. Thirdly, we use n-gram studies to examine where there are common ground areas of lexical and conceptual overlap and similarity between the two subcorpora, as well as similarities within the subcorpora indicative of copy-paste responses. Overall, these three analyses – unique lexis demonstrating strong oppositionality, semantic domains displaying changes in conceptual focus, and n-grams of overlap showing common ground –provide a fertile space for the exploration of the discourse of polarised and contentious issues in healthcare.

References

Health, Social Care and Sport Committee (2024), *CitizenSpace Call for Views on Assisted Dying for Terminally Ill Adults,* Edinburgh: Scottish Parliament.

McArthur, Liam (2022), *Proposed Assisted Dying for Terminally Ill Adults Bill: Consultation Document,* Edinburgh: Scottish Parliament.

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

Rayson, Paul (2009), *Wmatrix: a web-based corpus processing environment,* Lancaster: Lancaster University. Available at: http://ucrel.lancs.ac.uk/wmatrix/

Automatic identification and functional classification of multi-word expressions for diachronic analysis of scientific English

Diego Alves, Stefan Fischer and Elke Teich
(Saarland University)

In this study, we examine multi-word expressions (MWEs) and their functions in scientific writing, exploring their diachronic changes from the mid-17th century to the present. MWEs contribute to language efficiency by serving as highly predictable linguistic elements that offer a clear processing advantage. Their role in scientific writing is especially noteworthy given the substantial informational density within the scientific field, where MWEs can help smooth the cognitive load for a more efficient flow of information (Conklin and Schmitt, 2012). Alves et al. (2024) identified and characterized 552 MWEs with functional labels to examine changes in their usage over time in the Royal Society Corpus (RSC; Fischer et al., 2020), a diachronic corpus of scientific English spanning the period from 1665 to 1996, based on the Philosophical Transactions and Proceedings of the Royal Society of London and comprising 47,837 texts (295,895,749 tokens), primarily scientific articles across a broad range of disciplines, including the mathematical, physical, and biological sciences. The functions of stance expressions (e.g., *it is important*, *needs to be*), discourse organizers (e.g., *in this article*, *due to the fact*), and referential expressions (e.g., *a form of*, *at the end of*) are derived from previous research grounded in Hallidayan register theory (Halliday and Matthiessen, 2014), also adopted by English for Academic Purpose scholars (Biber et al., 2004; Simpson-Vlach and Ellis, 2010; Liu, 2012). Their results show, for example, an increase in the typicality of stance expressions in the 20th century. However, due to some limitations in the methods, only a restricted number of MWEs was considered.

Therefore, we complemented the approach presented by Alves et al. (2024) to identify and classify MWEs using embedding spaces (per decade of the RSC) built with structured skip-grams (Ling et al., 2015). By identifying the closest neighbors of the previously classified MWEs in the embedding spaces, followed by a manual verification, we were able to extend the list with other expressions that were neglected in the previous study, achieving a total of 807 MWEs (i.e., 201 discourse organizers, 412 referential expressions, and 194 stance expressions). This increase of 46% shows that the embedding approach is a useful way of improving the identification and classification of MWEs, especially regarding diachronic data as it considers all periods individually.

With this improved set of MWEs, the idea is to identify evolutionary trends using relative entropy, specifically Kullback-Leibler Divergence (KLD; Kullback and Leibler, 1951), which is a method for comparing probability distributions. KLD measures the additional number of bits required to encode a given data set A when a non-optimal model based on data set B is applied to a set of elements X (each MWE functional class in this study). We expect to confirm the general trends presented previously with a more precise view of the changes regarding MWE usage over time, also identifying the MWEs of each class contributing the most to the observed tendencies.

References

Alves, Diego, Stefania Degaetano-Ortlieb, Elena Schmidt, and Elke Teich (2024), 'Diachronic Analysis of Multi-Word Expression Functional Categories in Scientific English', in *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, 81–87.

Biber, Douglas, Susan Conrad, and Viviana Cortes (2004), If you look at…: Lexical bundles in university teaching and textbooks, *Applied Linguistics* 25(3), 371–405.

Conklin, Kathy, and Norbert Schmitt (2012), The processing of formulaic language, *Annual Review of Applied Linguistics* 32, 45–61.

Fischer, Stefan, Jörg Knappen, Katrin Menzel, and Elke Teich (2020), 'The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study', in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 794–802.

Halliday, Michael Alexander Kirkwood, and Christian M. I. M. Matthiessen (2013), *Halliday's Introduction to Functional Grammar*, London: Routledge.

Kullback, Solomon, and Richard Allen Leibler (1951), 'On Information and Sufficiency', *The Annals of Mathematical Statistics* 22(1), 79–86. https://doi.org/10.1214/aoms/1177729694.

Ling, Wang, Chris Dyer, Alan W. Black, and Isabel Trancoso (2015), 'Two/Too Simple Adaptations of Word2Vec for Syntax Problems', in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2015)*, Denver, CO, 1299–1304.

Liu, Dilin (2012), 'The Most Frequently-Used Multi-Word Constructions in Academic Written English: A Multi-Corpus Study', *English for Specific Purposes* 31(1), 25–35.

Simpson-Vlach, Rita, and Nick C. Ellis (2010), 'An Academic Formulas List: New Methods in Phraseology Research', *Applied Linguistics* 31(4), 487–512.

Does text simplification affect the use of multi-word expressions? A corpus-based analysis of English biomedical abstracts

Sergei Bagdasarov, Elke Teich and Diego Alves
(Saarland University)

Work-In-Progress

We focus on multi-word expressions (MWEs), prefabricated formulaic sequences of words that constitute large proportions of language and contribute to language fluency [5], in complex and plain English biomedical abstracts. Starting from the premise that language production varies depending on the situational context of communication [3], we assume that the change in target audience background (specialist-to-specialist vs specialist-to-layperson communication) will be reflected in the use of MWEs in complex and plain abstracts. Accordingly, our research question is: How does the shift in complexity affect MWE usage?

Our analysis is based on the Plain Language Adaptation of Biomedical Abstracts (PLABA) dataset [2], which is a parallel corpus of biomedical abstracts and their human-created plain English adaptations. Following [1], our approach to MWE identification is threefold. First, we rely on the Universal Dependencies framework [9] that contains syntactic relations commonly associated with formulaic language: **compound** (*chain reaction*), **compound:prt** (phrasal verbs: *figure out*), **fixed** (grammaticalized expressions: *due to*), **flat** (proper names: *Moderna mRNA-1273*). We parsed the corpus with the state-of-the-art stanza parser [10] and extracted all words tagged with these labels and their corresponding heads. Second, we use the Academic Formulas List, short: **AFL** [11], which is a set of expressions characteristic of academic language (*large number of*, *keep in mind*). This list was compiled based on a measure called formula teaching worth (FTW) – a combination of frequency and mutual information. And finally, we use Partitioner [12] – a Python module for MWE extraction that employs a supervised machine learning algorithm [13].

For the comparative analysis of plain vs. complex abstracts, we employed the asymmetric variant of relative entropy, also known as Kullback-Leibler Divergence (KLD) [6]. KLD is an information-theoretic measure that allows us to quantify the difference (in bits) between two probability distributions (in our case, MWEs in complex and plain abstracts) and identify features contributing to the divergence. The benefit of KLD lies in uniting frequency and typicality in one measure, which we will explain in detail in our talk (see [7,8]). Also, it allows us to capture differences in both directions of comparison.

In addition to term patterns, we specifically analyzed discourse-related MWEs (**fixed** and **AFL**) and their functionality. For this, we manually classified the MWEs using the taxonomy proposed by [4], which includes three major groups: **stance expressions**, **discourse organizers** and **referential expressions**. These are further divided into more fine-grained subcategories. Our analysis revealed that the most notorious trend concerns terminology patterns which undergo different manipulations in plain abstracts: e.g.

omission of statistical and methodological terminology (*confidence interval*), terminology transformations (*brain disorder* instead of *neurodegenerative disease*). Furthermore, MWEs used for topic introduction (*we looked at*) and focus identification (*that is*) are more characteristic of plain abstracts. Different referential expressions specifying attributes are common both in complex and plain abstracts, while the latter mostly feature MWEs referring to quantities or tangible attributes (*the amount of*, *the size of*) and the former are more characterized by MWEs framing intangible attributes (*on the basis of*).

References

Alves, Diego, Stefania Degaetano-Ortlieb, Elena Schmidt, and Elke Teich (2024), Diachronic analysis of multi-word expression functional categories in scientific English, in *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, Torino, Italia: ELRA and ICCL, 81–87.

Attal, Kelly, Brian Ondov, and Dina Demner-Fushman (2023), A dataset for plain language adaptation of biomedical abstracts, *Scientific Data* 10(1):8.

Biber, Douglas (2012), Register as a predictor of linguistic variation, *Corpus Linguistics and Linguistic Theory* 8(1): 9–37.

Biber, Douglas, Susan Conrad, and Viviana Cortes (2004), If you look at... : Lexical bundles in university teaching and textbooks, *Applied Linguistics* 25(3):371–405.

Conklin, Kathy, and Norbert Schmitt (2008), Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers?, *Applied Linguistics* 29(1):72–89.

Fankhauser, Peter, Jörg Knappen, and Elke Teich (2014), Exploring and visualizing variation in language resources, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland: ELRA, 4125–4128.

Gries, Stefan Th. (2024), *Frequency, Dispersion, Association, and Keyness: Revising and Tupleizing Corpus-Linguistic Measures*, Amsterdam & Philadelphia: John Benjamins.

Kullback, Solomon, and Richard A. Leibler (1951), On information and sufficiency, *The Annals of Mathematical Statistics* 22(1):79–86.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman (2021), Universal dependencies, *Computational Linguistics* 47(2):255–308.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (2020), Stanza: A Python natural language processing toolkit for many human languages, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online: Association for Computational Linguistics, 101–108.

Simpson-Vlach, Rita, and Nick C. Ellis (2010), An academic formulas list: New methods in phraseology research, *Applied Linguistics* 31(4):487–512.

Tanner, Joshua, and Jacob Hoffman (2023), MWE as WSD: Solving multiword expression identification with word sense disambiguation, in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore: Association for Computational Linguistics, 181–193.

Williams, Jake (2017), Boundary-based MWE segmentation with text partitioning, in *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Copenhagen, Denmark: Association for Computational Linguistics, 1–10.

Building the Corpus of American Style Guides (CASG): A tool for linguistic research

Holly Baker
(Brigham Young University)

Linguistic prescriptivism is an emerging discipline in the field of linguistics (Beal, Luckac, & Straaijer, 2023) and often involves corpus-based approaches to research (Szmrecsanyi & Bloemen, 2023). Although much work has been done in building corpora of English language texts (e.g., COCA, OEC,

TIME), creators have done little to acknowledge that many texts included in specialized corpora have been subjected to a prescriptive editing process. As a prescriptive tool, style guides have served as an authoritative and sometimes dogmatic influence on editorial decision-making since the early 20th century and continue to impact editorial practices today. Style guides such as *The Chicago Manual of Style* (CMOS), the *Associate Press Style Guide*, and others have increasingly included prescriptions on grammar, usage, and stylistic prose. In studying the work of editors and the role of style guides, however, we should keep in mind what Linda Pilliere (2020) has said about that relationship: "How far copy editors reflect the values of style and usage guides in their decisions … is difficult to measure." Nevertheless, having a database of style guides may enable researchers to ask and answer questions they were not able to before in the absence of such a resource.

The groundwork for research into style guides and their impact on the editorial process and the language is already being laid (Straaijer 2015, Tieken-Boon van Ostade 2020). However, currently, no database or corpus of style guides exists. Some researchers (see, for example, Chang & Swales, 1999; Bennett, 2009) have compiled their own limited corpora of style guides for the purposes of research, but these corpora are neither exhaustive nor public. As studies in linguistic prescriptivism (Curzan, 2014; Klein, 2005; Pullam, 2004) and editing and publishing (Flanagan, 2019; Lang & Palmer, 2017 Melonçon, 2019) become more prevalent among scholars, a more comprehensive and up-to-date database will be necessary in facilitating data collection and analysis of the role of style guides in language studies broadly. Therefore, just as Straaijer et al. have created a database of usage guides (HUGE: The Hyper Usage Guide of English), I have undertaken to create a corpus of style guides—the Corpus of American Style Guides (CASG)— to aid in answering the broad question: How do style guides contribute or respond to the enforcement of norms across time? In this work-in-progress report, I will introduce the English corpus community to the CASG as a robust database of American style guides spanning the last century to the present, to include academic, trade, corporate/news media, and government/NGO style guides. I will report on its progress and usability, including search features (e.g., KWIK, concordance lines) and results data (e.g., raw number and normalized frequency, keyword analysis, collocations), and filters (e.g., edition, genre, year range). Finally, I will announce when the CASG is expected to be publicly available to scholars to aid in their own research, as it is intended to assist scholars researching in fields such as linguistics (including corpus linguistics, prescriptivism, etc.), editing, publishing, technical/business communication, science communication, and education.

References

Beal, Joan C., Lukac, Morana, & Straaijer, Robin. (2023). *The Routledge handbook of linguistic prescriptivism*. Routledge.

Bennett, Karen. (2009). English academic style manuals: A survey. *Journal of English for Academic Purposes*, 8(1), 43–54.

Chang, Y., & Swales, J. (1999). Informal elements in English academic writing: Threats or opportunities for advanced non-native speakers? In C. Candlin, & K. Hyland (Eds.), Writing: Texts, processes and practices (pp. 145e167). Essex, UK: Addison Wesley Longman Ltd.

Curzan, Anne. (2014). *Fixing English: Prescriptivism and language history*. Cambridge, UK: Cambridge Univ. Press.

Flanagan, Suzan. (2019). The current state of technical editing research and the open questions. In Flanagan, Suzan, & Albers, Michael J. (Eds.)*, Editing in the modern classroom* (pp. 15–46). Routledge.

Klein, Wolf Peter. (2005). Deskriptive statt Präskriptiver Sprachwissenschaft!? *Zeitschrift für Germanistische Linguistik* 32.3: 376–405.

Lang, Susan, & Palmer, Laura. (2017). Reconceiving technical editing competencies for the 21st century: Reconciling employer needs with curricular mandates. *Technical Communication, 64*(4), 297–309. https://www.jstor.org/stable/26464505

Melonçon, Lisa. (2019). *Editing in the modern classroom* (Eds. Suzan Flanagan & Michael J. Albers). Taylor & Francis.

Pillière, Linda. (2020). US copy editors, style guides and usage guides and their impact on British novels. *Language Prescription*. Multilingual Matters.

Pullum, Geoffrey K. (2004). Ideology, power and linguistic theory. Paper presented at the Annual Meeting of the Modern Language Association, Philadelphia, 30 December 2004.

Straaijer, Robin. (2015). The Hyper Usage Guide of English (HUGE) database User Manual. *version 20150218*.

Szmrecsanyi, Benedikt, & Bloemen, Dieuwrtje. (2023). Corpus-based approaches to prescriptivism. *The Routledge Handbook of Linguistic Prescriptivism* (eds. Beal, Lukac, and Straaijer). Routledge.

Tieken-Boon van Ostade, Ingrid. (2020). *Describing prescriptivism: Usage guides and usage problems in British and American English*. Routledge.

Beware of the Hoover: a corpus-driven investigation of representations of borderline personality disorder (BPD) on the *r/BPDlovedones* subreddit

James Balfour
(University of Glasgow)

This study examines how friends, family members, and loved ones of people with Borderline Personality Disorder (BPD) linguistically construct and understand the disorder on *r/BPDlovedones*, a Reddit community dedicated to discussing experiences with BPD-diagnosed individuals. This investigation is particularly urgent given that BPD affects approximately 2% of the UK population (1.3 million people) (McManus et al, 2014), with those diagnosed being especially vulnerable to stigma and negative outcomes (e.g. Kling, 2014; Soderholm et al, 2020).

The research employs corpus-driven discourse analysis to examine linguistic patterns and terminology in user discussions. This approach is situated within a growing body of research investigating illness construction in online settings (e.g. Hunt and Brookes, 2020), with particular attention to how intimate relationships with BPD individuals are conceptualized and negotiated through language. The study is uniquely positioned to examine how anonymous online spaces facilitate discussion of topics that might be considered taboo or stigmatizing in other contexts, particularly focusing on how personal experiences shape broader narratives about the disorder.

The study analyzes posts from *r/BPDlovedones*, a subreddit with 6.81k members that describes itself as "a safe space for people to discuss the challenges and abuse they have endured at the hands of someone who has Borderline Personality Disorder." The corpus comprises all comments posted between January 2011 and December 2019, totalling 29 million words. This represents one of the largest specialized corpora of online forum data focused on a specific mental health condition.

The analysis compares the *r/BPDlovedones* corpus against a reference corpus (*the Reddit Conversation Corpus*, containing 69,428,488 words from 95 random subreddits). Keywords were identified using Log Ratio, with analysis focusing on the top 100 words. All selected keywords had log-likelihood scores above 6.63, ensuring a 99% confidence level that frequency differences were not due to random variation.

The analysis reveals that users develop and employ site-specific jargon, particularly acronyms like PWBPD (person with BPD), EXBPD (ex-partner with BPD), and notably UBPD (undiagnosed BPD), which are problematically used to label individuals without formal diagnoses. The of specialized terminology suggests the development of an in-group discourse that may both support and stigmatize. While some terminology aligns with formal diagnostic criteria (e.g., "abandonment"), many site-specific terms (e.g., "mirroring," "discard," "hoover") have emerged to describe perceived behavioural patterns. The linguistic evolution of terms like "hoover" from verbs (“she hoovered me”) to nouns (“a hoover”) demonstrates how individual experiences become reified as clinical categories within the community. Analysis of collocates around words like "self" and "person" reveals conflicting narratives about the "authentic" identity of individuals with BPD, with some users attributing authenticity to idealization phases and others to devaluation phases.

These competing narratives highlight the complexity of how BPD is understood and constructed by those closely affected by it.

The findings suggest that while online communities provide valuable support spaces, they can also propagate potentially problematic lay diagnoses and stereotypes about BPD through specific linguistic choices and patterns. This has implications for both mental health professionals and online community moderators in understanding how informal support spaces might influence perceptions and treatment of BPD.

References

Hunt, Daniel, and Gavin Brookes (2020), *Corpus, Discourse and Mental Health*, London: Bloomsbury Academic.

Kling, Rachel (2014), Borderline Personality Disorder, language, and stigma, *Ethical Human Psychology and Psychiatry* 16.2: 114–119. https://doi.org/10.1891/1559-4343.16.2.114

McManus, Sally, Paul E. Bebbington, Rachel Jenkins, and Traolach S. Brugha (eds) (2016), *Mental Health and Wellbeing in England: Adult Psychiatric Morbidity Survey 2014*, Leeds: NHS Digital.

Söderholm, John J., J. Lumikukka Socada, Tom Rosenström, Jesper Ekelund, and Erkki T. Isometsä (2020), Borderline Personality Disorder with depression confers significant risk of suicidal behavior in mood disorder patients – A comparative study, *Frontiers in Psychiatry* 11: 290. https://doi.org/10.3389/fpsyt.2020.00290

Bipartite collexeme graphs for the diachronic analysis of constructional competition

Andreas Baumann, Axel Bohmann and Lotte Sommerer
(University of Vienna, University of Cologne, FAU Erlangen-Nürnberg)

In this talk, we develop a method for tracing language change, conceptualized as competition among constructions, from a collostructional perspective (Stefanowitsch & Gries 2003). Specifically, we extend the network-based methodology introduced in Sommerer and Baumann (2021) to incorporate a diachronic dimension. To do so, we analyze the collostructional profiles of future-time referring (FTR) constructions across 200 years of American English, as represented in the Corpus of Historical American English (COHA; Davies 2010-). We construct what we refer to as bipartite collexeme graphs to model collostructional relationships. In a nutshell, such a network graph consists of nodes for verbs (like *eat* or *die*) as well as FTR constructions (*will, shall, be going to*); edges connecting verbs with constructions represent the strength of association between them. For any pair of verb and FTR construction, we test whether the two members are sufficiently strongly attracted to each other in the sense that the verb occurs in the construction (i) with an above-chance probability, and (ii) to a substantial degree. Both i) and ii) are derived from a corpus-based contingency table displaying frequencies A-D:

	[*will* V]	not [*will* V]
eat	A = frequency of [*will eat*]	B = frequency of *eat* outside of [*will* V]
not *eat*	C = frequency of [*will* V] without V being *eat*	D = frequency of all other relevant constructions (here: all non-FTR verb phrases)

We construct network representations of the kind outlined above for ten intervals of 20 years in COHA. On the basis of these, we are able to a) trace the overall size of the network associated with each construction, b) the specific verbs associated with each, and c) the change in association patterns in particular words and in aggregate (see Figure 1).

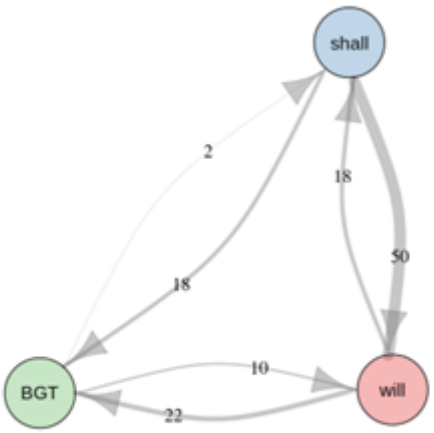


Figure 1: Aggregate flow analysis of verbs changing their association with FTR constructions over time.

Our findings confirm previous research, both in demonstrating how the overall rise in frequency of *be going to* is accompanied by an expansion of the network of verbs associated with this construction as well as by finding previously attested semantic differentiation between verbs associated with *will* and *be going to* (Gries & Stefanowitsch 2004). However, the method also generates new findings, such that *be going to* becomes increasingly associated with concrete verbs (Brysbaert et al. 2014) over time instead of semantically bleaching, as one might expect as an outcome of ongoing grammaticalization.

References

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman (2014), Concreteness ratings for 40 thousand generally known English word lemmas, *Behavior Research Methods* 46, 904–911.

Davies, Mark (2010-), *The Corpus of Historical American English.* http://corpus.byu.edu/coha/

Gries, Stefan Th. and Anatol Stefanowitsch (2004), Extending collostructional analysis: A corpus-based perspective on 'alternations', *International Journal of Corpus Linguistics* 9(1), 97–129.

Sommerer, Lotte and Andreas Baumann (2021), Of absent mothers, strong sisters and peculiar daughters: The constructional network of English NPN constructions, *Cognitive Linguistics* 32(1), 97–131.

Stefanowitsch, Anatol and Stefan Th. Gries (2003), Collostructions: Investigating the interaction between words and constructions, *International Journal of Corpus Linguistics* 8(2), 209–243.

If only we would combine corpora and teaching: A qualitative analysis of if-clauses in the Corpus of Young German Learner English

Lea Bracke
(University of Bamberg)

Work-In-Progress

Ever since the emergence of Learner Corpus Research (LCR) as a distinguishable research area in the 1990s (Tracy-Ventura & Paquot 2020: 3), it has extended its scope both methodologically and content-wise. Despite its brief past, it can be argued that "the future of learner corpus research looks distinctly promising" (Huat 2013: 202). As an interdisciplinary field (Granger et al. 2015: 3), LCR holds valuable insights for both corpus linguistics and EFL teaching.

Due to the overrepresentation of written and cross-sectional data from advanced learners, a key future goal of LCR is to transform this status-quo by gathering other types of data (Tracy-Ventura & Paquot 2021: 4). The *Corpus of Young German Learner English* (YGLE; Werner et al. in preparation) attempts to address some of these research gaps by:

i. collecting data from beginner to intermediate learners of English;
ii. including a variety of task types across different modalities (spontaneous speech; short essay tasks, and a digital communication task);
iii. gathering pseudo-longitudinal data across eight grade levels

To illustrate the opportunities offered by such data, this case study will analyse the use of *if*- clauses in YGLE across different age cohorts.

Despite being taught rigidly in German secondary schools – usually as conditionals type 0-III (Endley 2010: 442) – students frequently struggle with the form, meaning, and time-tense relationship of *if*-clauses (Römer 2007: 360; Celce-Murcia and Larsen-Freeman 1999: 340; see also Covitt 1976). Simultaneously, *if*-clauses lend themselves well to the investigation of error gravity (cf. Evans et al. 2014), since not all deviations from the *if*-clause types commonly taught in German secondary schools are equally severe. The wording *If I would*, for instance, may be considered erroneous from a prescriptive point of view, but it frequently occurs in colloquial native speech (Celce-Murcia and Larsen-Freeman 1999: 344). Hence, a nuanced perspective on *if*-clauses and a focus on individualisation – i.e. the adaptation of tasks, instructions, and learning goals to individual students (Bray & McClaskey 2013: 1) should arguably be fostered in EFL teaching. This study therefore follows a descriptive approach to accuracy, investigating the following guiding research question: Which difficulties do beginner to intermediate learners of English encounter when using *if*-clauses?

To this end, a qualitative analysis of the Bavarian sub-sample of the YGLE corpus (*N* = 800 students) will be conducted. The results will be manually coded and categorized into i) formal errors (*If I would…*); ii) incoherence of modality or tense (**If I had …, I will …*); iii) inept verbal combinations (**If I had win*…), and iv) incorrect use of conjunctions (e.g. *when* instead of *if*).

Subsequently, some suggestions for addressing the actually observed problem areas in EFL education will be provided. The focus hereby will lie on how the insights gained from YGLE may inform error correction tasks, and activities fostering data-driven learning (DDL; sensu Johns 1991). Eventually, these tasks will be adjusted to various levels of difficulty and to different learning objectives, keeping the concept of individualization in mind.

References

Bray, Barbara and Kathleen McClaskey (2013), "Personalization v differentiation v individualization chart (v3)". https://www.marshfieldschools.org/cms/lib/WI01919828/Centricity/Domain/82/PL_Diff_Indiv.pdf (11 January 2025).

Celce-Murcia, Marianne and Diane Larsen-Freeman (1999), *The grammar book: An ESL/EFL teacher's course (2nd edition),* New York: Heinle and Heinle.

Covitt, Regina I. (1976), *Some problematic grammar areas for ESL teachers,* Dissertation, University of California, Los Angeles

Endley, Martin J. (2010), *Linguistic perspectives on English grammar: A guide for EFL teachers,* Charlotte, NC: Information Age Publishing.

Evans, Norman W., K. James Hartshorn, Troy L. Cox, and Teresa M. De Jel (2014), Measuring written linguistic accuracy with weighted clause ratios: A question of validity, *Journal of Second Language Writing* 24, 33–50.

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier (2015), Introduction: learner corpus research – past, present and future, in S. Granger, G. Gilquin, and F. Meunier (eds), *The Cambridge handbook of learner corpus research,* Cambridge University Press, 1–5.

Huat, Chau M. (2013), Learner Corpora and Second Language Acquisition, in K. Hyland, C. M. Huat, and M. Handford (eds), *Corpus applications in applied linguistics,* London/New York: Bloomsbury, 191–207.

Johns, Tim F. (1991), Should you be persuaded: Two samples of data-driven learning materials, *English Language Research Journal* 4, 1–16.

Römer, Ute (2007), Learner language and the norms in native corpora and EFL teaching materials: A case study of English conditionals, in S. Volk-Birke and J. Lippert (eds), *Proceedings of the Conference of the German Association of University Teachers of English,* Trier: Wissenschaftlicher Verlag Trier, 355–364.

Tracy-Ventura, Nicole, and Magali Paquot (2020), Second language acquisition and corpora: An Overview, in N. Tracy-Ventura and M. Paquot (eds), *The Routledge handbook of second language acquisition and corpora,* New York: Routledge, 1–8.

Werner, Valentin, Robert Fuchs, Anna Rosen, Lyudmila Kruhlenko, Bethany Stoddard, and Lea Bracke, In preparation, *Corpus of Young German Learner English* (unpublished).

Linguistic variation between human-written and machine-generated text

Sophia Conrad
(University of Zurich)

Work-In-Progress

The rapid advancement of Large Language Models (LLMs) such as GPT has enabled the generation of texts that are grammatical and fluent, and it be- comes increasingly harder to distinguish machine-generated texts (MGT) from human-written texts (HWT). Currently, there are many studies that focus on the detection of MGT for plagiarism detection (see [10] for a review). Additionally, some work has analyzed the linguistic and stylistic differences between HWT and MGT. Previous work suggests that there are significant differences between HWT and MGT regarding grammatical, lexical, and stylistic features. However, most studies focus on a rather limited feature set (e.g. [7]) or on variation within a specific register (e.g. scientific texts [6], dialogues [8], or abstracts of research articles [9]). The goal of this work-in-progress study is to perform a thorough linguistic analysis of MGT to determine how closely it aligns with HWT across a broad range of registers. The **main research question** thus is: Can LLMs accurately replicate the distinct registers of human writing?

To address this question, I am using Biber's multidimensional analysis (MDA) framework [3], which has been widely used to study register variation (see [1] for a compilation of register studies employing MDA).

As a first step, a dataset of HWT and MGT is created. HWT samples are taken from the Corpus of Contemporary American English (COCA), a large and widely used corpus of one billion words with texts from eight different registers. Importantly, the most recent COCA texts date back to 2019, ensuring that no MGT is present in the dataset. Comparable MGT samples are generated using three popular LLMs (GPT-3.5, Llama3.3, and Falcon) under two different conditions: (1) prompting the models with the first sentence of the HWT, along with its register and desired length, and (2) providing half of a HWT and instructing the models to complete it with a certain length restriction. This allows to approach two **secondary research questions**: Can LLMs better replicate human written registers when given a longer example? And are there differences between the LLMs?

The next step is the comparison of linguistic features in the two text types using Biber's MDA. To that end, the MFTE tagger [5] is employed, which implements over 100 linguistic features described by [4]. The study identifies key dimensions of linguistic variation, such as information content, narrative spectrum, and context dependency, and examines whether MGT exhibits statistically significant differences from HWT across these dimensions.

Lastly, the results can be compared to the findings of previous work. Especially relevant will be the comparison to [2] because of the similar approach (MDA), the same model (GPT), but adding additional ones, and investigating different registers. I expect to obtain results that are in line with previous findings,

namely that there are significant differences between HWT and MGT and smaller differences within the variation of different LLMs.

References

Biber, Douglas (1991), *Variation across speech and writing,* Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999), *Longman Grammar of Spoken and Written English,* Harlow: Pearson Education Limited.

Federica Barbieri and Stacey Wizner (2019), Appendix A: Annotations of major register and genre studies, in D. Biber and S. Conrad, *Register, Genre, and Style*, 364.

Le Foll, Elen (2024), *Textbook English: A multi-dimensional approach,* John Benjamins Pub Co.

Markey, Ben, David West Brown, Michael Laudenbach, and Alan Kohler (2024), Dense and Disconnected: Analyzing the Sedimented Style of ChatGPT-Generated Text at Scale, *Written Communication* 41(4), 571-600. DOI: 10.1177/07410883241263528. URL: https://journals.sagepub.com/doi/10.1177/07410883241263528 (visited on 11/18/2024).

Rosenfeld, Ariel and Teddy Lazebnik (2024), *Whose LLM is it Anyway? Linguistic Comparison and LLM Attribution for GPT-3.5, GPT-4 and Bard,* ArXiv, abs/2402.14533. URL: http://arxiv.org/abs/2402.14533 (visited on 11/18/2024).

Sandler, Morgan, Hyesun Choung, Arun Ross, and Prabu David (2024), A Linguistic Comparison between Human and ChatGPT- Generated Conversations, in C. Wallraven, CL. Liu, A. Ross (eds), *Pattern Recognition and Artificial Intelligence,* ICPRAI 2024, Lecture Notes in Computer Science, vol 14892, Springer, Singapore. URL: http: //arxiv.org/abs/2401.16587 (visited on 11/14/2024).

Sardinha, Tony Berber (2024), AI-generated vs human-authored texts: A multidimensional comparison, *Applied Corpus Linguistics* 4(1), 100083. DOI: 10.1016/j.acorp.2023.100083. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666799123000436 (visited on 12/05/2024).

Theocharopoulos, Panagiotis C., Panagiotis Anagnostou, Anastasia Tsoukala, Spiros V. Georgakopoulos, Sotiris K. Tasoulis, and Vassilis P. Plagianakos (2023), Detection of Fake Generated Scientific Abstracts, in *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService),* Athens, Greece: IEEE, 33–39. DOI: 10.1109/BigDataService58306.2023.00011. URL: https://ieeexplore.ieee. org/document/10233982/ (visited on 12/16/2024).

Valiaiev, Dmytro (2024), Detection of Machine-Generated Text: Literature Survey, in *arXiv preprint arXiv:2402.01642.*

(T)here is an existential Ø misses its subject relativizer: On the history of zero subject relativizers in EModE

Robert Daugs and Marco Wiemann
(Kiel University)

This paper concerns zero subject relativizers (ZSRs) as in *there is no Fortress Ø can save him* (EEBO A01615). ZSRs were the main type of non-marked relative constructions in Old and Middle English (Fischer & van der Wurff 2008: 128), while in present-day standard varieties examples of zero object relativizers are much more common as in *this is the book Ø he was talking about* (BNC F8R 1001). Yet, ZSRs are still attested today in traditional dialects (Beal & Corrigan 2005: 212-3) and in informal settings and spoken registers (Lehmann 2002, Kaltenböck 2023). Nowadays, they can mainly be found with locative *here*, existential *there* and in *it*-cleft constructions, although other types exist (Lehmann 2002: 172). Using the *Corpus of English Dialogues*, Tottie and Johansson (2015) show a downward trend of ZSRs from the beginning of the 17th century to the 18th century and find that they are more frequent in comedies than trials.

However, comprehensive historical studies with corpora larger than 10 million words are still wanting. Therefore, we investigate how ZSRs developed from 1600 until the beginning of the 18th century in *Early English Books Online V3* (EEBO), a corpus which contains ca. 1.2 billion words. We analyse to what extent trends identified by Tottie and Johansson (2015) are mirrored in EEBO and which factors drive the development of ZSRs. We consider factors such as genre/register, time, and prosody. To assess their effect on the choice between ZSRs and related constructions with expressed pronouns *who*, *which*, and *that*, the data were submitted to generalized linear mixed modelling (Gries 2021).

In contrast to Tottie and Johansson (2015), we show that *(t)here* ZSRs show a non- monotonous trend rather than a monotonous downward trend. Furthermore, *it*-cleft ZSRs in fact show a significant increase throughout the 17th century. In terms of genre, plays, and poetry and song are the best predictors of ZSRs in our data. We argue that this could be explained by their speech-like (and speech purposed) characteristics (Culpeper & Kytö 2010). Moreover, ZSRs could be utilized in these genres to fit the required metre, i.e., mostly iambic pentameter, which was a crucial feature of dialogues in EModE plays, as illustrated for instance by these lines from Shakespeare's *Henry VIII* (1613): *And Corne shall flye asunder. For I know There's none Ø stands vnder more calumnious tongues* (Act 5.Scene 1). We hypothesize that the prosody of utterances containing ZSRs may play a role beyond metre conventions required by certain genres.

References

Beal, Joan C. and Karen P. Corrigan (2005), A tale of two dialects: Relativization in Newcastle and Sheffield, in Markku Filppula, Juhani Klemola, Marjatta Palander, and Esa Penttilä (eds), *Dialects Across Borders Selected Papers from the 11th International Conference on Methods in Dialectology (Methods XI), Joensuu,* Amsterdam & Philadelphia: John Benjamins, 211–29.

BNC Consortium, *The British National Corpus, XML Edition (*2007), Oxford Text Archive. http://hdl.handle.net/20.500.14106/2554.

Culpeper, Jonathan and Merja Kytö (2010), *Early Modern English Dialogues: Spoken Interaction as Writing,* Cambridge: Cambridge University Press.

EEBO. *Early English Books Online (V3),* accessed through CQPweb, Lancaster University. Available online at https://cqpweb.lancs.ac.uk/eebov3/ accessed on 20 December 2024.

Fischer, Olga and Wim van der Wurff (2008), Syntax, in Richard Hogg and David Denison (eds), *A History of the English Language*, Cambridge: Cambridge University Press, 109–98.

Gries, Stefan Th. (2021), *Statistics for Linguistics with* R*: A practical Introduction*, 3rd ed., Berlin: De Gruyter.

Kaltenböck, Gunther (2023), On the use of *there*-clefts with zero subject relativizer, in Caroline Gentens, Lobke Ghesquière, William B. McGregor, and An Van linden (eds), *Reconnecting Form and Meaning: In Honour of Kristin Davidse*, Amsterdam & Philadelphia: John Benjamins, 17–43.

Lehmann, Hans Martin (2002), Zero subject relative constructions in American and British English, in Pam Peters, Peter Collins, and Adam Smith (eds), *New Frontiers of Corpus Research: Papers from the Twenty First International Conference on English Language Research on Computerized Corpora Sydney 2000*, Amsterdam & New York: Rodopi, 163–77.

Tottie, Gunnel and Christine Johansson (2015), *Here is an Old Mastiffe Bitch Ø Stands Barking at Mee*: zero subject relativizers in Early Modern English *(t)here-constructions,* in Peter Sundkvist, Philip Shaw, Britt Erman, and Gunnel Melchers (eds), *From Clerks to Corpora: Essays on the English Language Yesterday and Today*, Stockholm: Stockholm University Press, 135–53.

Propagation of structured variation across non-contiguous speaker groups: Intensifiers in FL English spoken in Germany

Julia Davydova
(University College of Teacher Education Vorarlberg)

An established wisdom of modern sociolinguistic discourse posits that the propagation of structured variation follows the path of interpersonal interactions amongst speakers (Labov 2001). Simultaneously, there is a growing sentiment amongst linguists that there is a possibility that patterns of variation may be spreading via channels other than those offered by (extensive) face-to-face communication amongst members of a speech community (Androutsopoulos 2014). This is because digital technologies have revolutionized the ways in which people interact with each other, making far-flung contacts instantly (and oftentimes) intimately available to each other. And also because recently developed video-content and online streaming platforms (*YouTube*, *Netffix*, etc.) has made a host of English lects and registers widely accessible to global audiences including countries of Western Europe (see *inter alia* Author 2019, 2025). Against this backdrop, new sociolinguistic models propose that the world-wide spread of sociolinguistic patterns of variation must necessarily involve systematic comparisons of the linguistic behaviour of (i) the donor communities, (ii) mass media texts and (iii) adopting speaker groups (Sayers 2014). This variationist study seeks to investigate (iii), while also keeping an eye on the patterns of sociolinguistic variation in (i) and (ii) as documented by the previous research (Author submitted; Reichelt C Durham (2017); Stratton (2020); Tagliamonte 2016). To that end, the investigation targeted the variable of intensifiers (*This is so cool*; *it's really amazing*; *I am very astonished*, etc.), while drawing on spontaneous speech data elicited from 36 EFL young adults (aged 18 to 26) from Germany, all of whom reported extensive exposure to English-language mass media texts through online streaming practices (*Netffix*, etc.). Performed through the script created in the software environment *R* using *glmer* function of the *lme4* package, the analyses of intensification pinpoint unambiguous similarities in the patterns of intensification attested in the German learner data on the one hand and mass media texts as well as donor (North American English) speech communities on the other. I discuss the implications that these findings carry for further building of sociolinguistic theory for a globally available language.

References

Androutsopoulos, Jannis (ed.) (2014), *Mediatization and Sociolinguistic Change*, Berlin: De Gruyter.

Davydova, Julia (2025), 'EFL adolescents' use of English in the era of new digital media: An empirical investigation', *International Journal of Applied Linguistics*, 35: 617–628.

Davydova, Julia (2019), *Quotation in Indigenised and Learner English: A Sociolinguistic Account of Variation*. Berlin & Boston: Mouton de Gruyter.

Labov, William (2001), *Principles of Linguistic Change, Volume 2: External Factors*, Oxford: Blackwell.

Reichelt, Susan, and Mercedes Durham (2017), 'Adjective Intensification as a Means of Characterization: Portraying In-Group Membership and Britishness in *Buffy the Vampire Slayer*', *Journal of English Linguistics* 45(1), 60–87. https://doi.org/10.1177/0075424216669747

Sayers, Dave (2014), 'The Mediated Innovation Model: A Framework for Researching Media Influence in Language Change', *Journal of Sociolinguistics* 18(2), 185–212.

Stratton, James M. (2020), 'Fiction as a Source of Linguistic Data: Evidence from Television Drama', *Token: A Journal of English Linguistics*, 10, 39–58.

Language and space exploration: Talking the future of humanity

Daria Dayter
(Tampere University)

Work-In-Progress

This study examines the linguistic representation of space exploration in Anglophone news discourse, aiming to understand how social actors, institutions, and the concept of space travel are constructed through language. For this study we compiled a corpus of approx. 3,000,000 words compiled from NexisUni, using news articles published from the 1960s to the present. Articles are retrieved using the search term *astronaut*, focusing on Anglophone sources from the regions North America, Europe, Australia & Oceania. This dataset is then investigated using a corpus-based analysis for a general linguistic description and a corpus-assisted discourse analysis (CADS) for a more in-depth study of discourse patterns.

The study addresses three central questions: How are social actors (i.e., astronauts) represented linguistically in news discourse? How do the representations social actors compare in the news from different regions and across time? What linguistic features characterize the discourse around space travel, particularly in terms of probability markers and thematic framing?

Methodologically, the study employs CADS techniques (Baker, 2006; Partington et al., 2013) to identify recurring lexical and grammatical patterns. Specific attention is given to social actor representation (using van Leeuwen's (2008) framework to explore agency, roles, exclusion, and specific versus generic descriptions of actors). We also focus on probability and framing by analysing modal verbs, adverbials, and evaluative language to identify three key discourses that present space travel as a realistic future endeavour, an aspirational goal, or an unattainable fantasy. Among other markers, this is reflected in the use of speculative and aspirational language (high-frequency markers of uncertainty "might," "could" and conditional constructions) in the latter two discourses. In contrast, the former discourse displays assertive certainty ("is set to," "will").

A preliminary study of the corpus as a whole suggests that news discourse predominantly frame space exploration as a symbol of technological progress and geopolitical competition, particularly during the Cold War era. Frame analysis has shown that the discourse emerging around ‘astronauts’ revolves around the professional trajectory and personal evolution. This discourse is seen emphasizing the prestige and dedication associated with the profession, as well as the institutional support and selection mechanisms in place. In addition, in analysing two periods (before and after 1995), the prominence of *communication* discourse remains strong in both periods, but the later period shows an increased emphasis on coming and visiting, suggesting more interaction and engagement. Further differences in emerging discourses reflect shifts in the roles and experiences of astronauts over time, highlighting changes in focus from purely operational tasks to more interactive and collaborative activities.

This study contributes to both linguistic and sociocultural understandings of space exploration, offering corpus-backed insights into how news media shape public perceptions.

References

Baker, Paul (2006), *Using Corpora in Discourse Analysis*, London: Continuum.

Partington, Alan, Alison Duguid, and Charlotte Taylor (2013), *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*, Amsterdam & Philadelphia: John Benjamins.

van Leeuwen, Theo (2008), *Discourse and Practice: New Tools for Critical Discourse Analysis*, Oxford: Oxford University Press.

A new dataset for email research and professional communication: Fauci2020

Rachele De Felice and Kate Warwick
(The Open University, WordSavvy)

As is well known, real-world datasets of email communication are very rare, due to commercial and personal sensitivities. Research has been – and continues to be – carried out on emails from the Enron corporation and the Hillary Clinton administration, but there is a need for more recent and varied data against which to verify those findings. The appearance of any new potential public email datasets is therefore a significant event in linguistics research. One likely new source are the over 2700 messages sent to and from Dr Anthony Fauci between January and June 2020 (the first six months of the COVID-19 pandemic), obtained by the Washington Post, Buzzfeed News, and CNN through Freedom of Information Act (FOIA) requests (see https://www.documentcloud.org/documents/20793561-leopold-nih-foia-anthony-fauci-emails). As one of the leading medical figures at the height of the pandemic, Fauci exchanged emails with a wide range of individuals including government officials, medical experts, the media, well-known individuals, and even the general public. This variety of interlocutors and the recency of the emails makes the dataset a promising new source of material for research on email communication, professional language, and politeness.

This presentation will introduce the Fauci dataset and explore its viability for corpus research, considering issues including formatting, redactions, and range of correspondents. It is currently available either as a single 3,234 page PDF, with redactions and repeated content, or as a JSON database (Benson et al. 2022). The latter provides a useful starting point as it also identifies all individuals exchanging emails, and groups emails into distinct threads. However, neither format is easily useable by most common corpus linguistics analysis programs. The presentation will discuss some of the tools to bridge the gap between these formats and the needs of corpus linguists.

It then presents an exploratory study of politeness norms in canonical requests, where findings can be easily compared to existing research. In particular, we ask: what differences are there between requests exchanged among very close collaborators (e.g. Fauci, his PA, and Chief of Staff) and those coming from complete strangers (ranging from fellow scientists and medical professionals to general members of the public)? The data was extracted using a combination of manual and automated searches, combining formal and functional approaches to speech act identification. Initial results show a surprising overall low frequency of indirect requests, and relatively more imperatives, often modulated by *please*. However, these appear to mainly cluster in exchanges among close collaborators, while exchanges with interlocutors outside the 'in-group' are characterised by higher degrees of formality, facework, and formulaicness.

Finally, the presentation will address the effects of familiarity and hierarchy on the formulation of requests, including a comparison with the Clinton Email dataset, which features similar ranges of interlocutors and a similar high-pressure work environment. This can contribute to identifying which features appear typical of (at least some) American English workplace requests. More generally, understanding the linguistic features of professional requests can be translated into actionable insights for businesses, showing how different linguistic choices can increase the success of their requests.

References

Benson, Austin R., Nate Veldt, and David F. Gleich (2022), 'fauci-email: A JSON Digest of Anthony Fauci's Released Emails', *Zenodo*. https://doi.org/10.5281/zenodo.5828209

Onomastic referencing in 18th-century British grammar writing

Nina Dumrukcic, Beatrix Busse and Sophie Du Bois
(University of Cologne)

The grammarians of the 18th century were well-known for their proclamations of what constitutes 'correct' language use, also known as the 'doctrine of correctness' (Leonard 1929/1962). For these, they appealed to the authority of analogy with Latin, reason, or logic. If we look at English grammar writing, Robert Lowth's *A Short Introduction to English Grammar* from 1762 and Lindley Murray's *The English Grammar* from 1795 are the most influential, and, probably, major prescriptivists (Tieken-Boon van Ostade 2011, p. 4) of the 18th century. Chapman (2008, p. 36) claims that the 18th-century grammarians were not solely prescriptivists, but that some investigated and published on language, attempting to address linguistic questions such as vernacular grammar or universals.

Within the HeidelGram project (https://heidelgram.de), investigations of British grammar writing from the 16th, 17th, and 19th centuries have indicated shifts in not only who is being referenced in these grammar texts, but also which strategies are employed to do so. This paper aims to continue the diachronic analyses of what we call onomastic, that is, name-based referencing by investigating who is referenced in 18th-century British grammars of English and in what way.

The HeidelGram corpus (Busse et al. 2015–) is a carefully designed corpus that comprises a representative selection of 16th- to 19th-century grammars of English. The 18th-century data comprises a total of about 1,5 million tokens taken from 24 grammar books. A citation network (see White 2011) of grammars and grammarians is created, which illustrates which texts from the 18th-century sub-corpus of the HeidelGram corpus refer to whom and who the most influential figures of the period are. Each onomastic reference is extracted automatically from corpus annotations using a custom tool written in Python and R in order to generate a network. The references are categorized into the seven person types that were established during the network analysis of the 16th-century grammar books (see Busse et al. 2021, 2024), such as *grammar author* and *ancient scholar*. Frequency and concordance analyses provide further quantitative and qualitative information which allow for a better understanding of these onomastic references.

It has been observed that the 18th-century "grammarians rely heavily on each other's work" (Locher 2008, p. 131), which suggests that there is significant referencing among the authors. We therefore expect to find higher proportions of references to grammar authors as compared to other person categories. References to works that are commonly considered to be prescriptive, such as Lowth's and Murray's, indicate how the authors in the corpus position themselves about the prescriptivism–descriptivism continuum. An ego network of references to Lowth will indicate his influence on the later 18th-century grammar authors.

References

Busse, Beatrix, Kirsten Gather, and Ingo Kleiber (2015), *HeidelGram. A Corpus of English Grammar Books between 1550 and 1900*. https://heidelgram.de

Busse, Beatrix, Nina Dumrukcic, and Sophie Du Bois (2024), *Onomastic Referencing Strategies in a Corpus of 17th-Century Grammars of English,* ICAME45, Vigo, Spain.

Busse, Beatrix, Nina Dumrukcic, Sophie Du Bois, and Ingo Kleiber (2021), A Corpus-Based Network Analysis of 16th-Century British Grammar Writing, *CL2021*, Limerick, Ireland.

Busse, Beatrix, Kirsten Gather, and Ingo Kleiber (2020), A Corpus-Based Analysis of Grammarians' References in 19th-Century British Grammars., in A. Cermakova, and M. Malá (eds.), Diskursmuster - Discourse Patterns: Vol. 20. *Variation in Time and Space: Observing the World Through Corpora*, De Gruyter.

Busse, Beatrix, Kirsten Gather, and Ingo Kleiber (2019), Paradigm Shifts in 19th-Century British Grammar Writing: A Network of Texts and Authors, in B. Bös, and C. Claridge (eds), *Norms and Conventions in the History of English*, John Benjamins.

Busse, Beatrix, Kirsten Gather, and Ingo Kleiber (2016), Assessing the Connections between English Grammarians of the Nineteenth Century: A Corpus-Based Network Analysis, in *Grammar and Corpora,* Heidelberg University Publishing, 435–442.

Chapman, Don (2008), The eighteenth-century grammarians as language experts, in I. Tieken-Boon van Ostade (ed.), *Grammars, Grammarians and Grammar-Writing in Eighteenth-Century England*, Berlin, New York: De Gruyter Mouton, 21–36.

Locher, Miriam A. (2008), Chapter 7: The Rise of Prescriptive Grammars on English in the 18th Century, in Fishman, Joshua A., Miriam A. Locher, and Jürg Strässler (eds), *Contributions to the Sociology of Language. Standards and Norms in the English Language* (95), Mouton de Gruyter, 127–148. https://doi.org/10.1515/9783110206982.1.127

Leonard, Sterling Andrus (1962), *The Doctrine of Correctness in English Usage 1700-1800,* Russel & Russel Inc. (Original work published 1929)

Tieken-Boon van Ostade, Ingrid (2008), Grammars, grammarians and grammar writing: An introduction, in I. Tieken-Boon van Ostade (ed), *Topics in English Linguistics: Vol. 59. Grammars, Grammarians and Grammar-Writing in Eighteenth-Century England*, Mouton de Gruyter, 1–14.

Tieken-Boon van Ostade, Ingrid (2011), *The bishop's grammar: Robert Lowth and the rise of prescriptivism in English,* Oxford: Oxford University Press.

White, Howard D. (2011), Scientific and scholarly networks, *The SAGE handbook of social network analysis*, 271–285.

A flexible approach to KWICGrouping concordances

Nathan Dykes, Stephanie Evert, Michaela Mahlberg and Alexander Piperski
(FAU Erlangen-Nürnberg)

Concordance analysis has a long-standing tradition as one of the fundamental techniques of corpus linguistics. It is supported by various corpus tools that offer different functionalities to help users organise concordance lines for interpretation. One such functionality is to select lines that contain specific words. In Sketch Engine (Kilgarriff et al. 2014), for example, users can filter concordances with regular expressions based on any annotation layer. KWICGrouper (O'Donnell 2008), which is currently implemented in the web app CLiC (Mahlberg et al. 2020), goes one step further: it allows users to specify multiple words which are then used to rank concordance lines according to the number of matches they contain.

In this contribution, we present a new implementation of KWICgrouper in FlexiConc, a dedicated Python library that offers a greater degree of control over parameter settings. Moreover, our library generates an analysis tree that stores all concordance views generated throughout the investigation to promote transparency by linking analytical insights to the corresponding concordance views. The library can be integrated into various host applications; for the present study, we use the web interface of CLiC.

To illustrate our approach, we explore patterns of body part nouns in English and German 19th-century novels. The English corpus is 19C of the CLiC corpora, and the German corpus has been compiled to be comparable to 19C (as explained in Finlayson et al. 2024). For each corpus, we create concordances for a range of common body part nouns (e.g. *foot, ears, cheek* and for English). We then use FlexiConc to analyse how body part nouns pattern in the two corpora. Using KWICGrouping operations with different settings, we highlight concordance lines with a high prevalence of particular linguistic features.

Settings implemented in our KWICGrouping framework include:

A. Token or type-based ranking: FlexiConc can rank lines based on either the count of distinct types or all token occurrences.

B. Regular expression support and case sensitivity control: this enables pattern-based searches with more fine-grained control, allowing the identification of morphological forms (e.g., *-ing* or *-ed*) or other linguistic features beyond specific words.

C. Context window customization: the start and end of the context window can be freely chosen to explore long-range dependencies.

D. KWICGrouping as filter: concordance lines satisfying certain KWICGrouping criteria can be selected and rearranged with different sorting, ranking, or grouping algorithms.

For instance, counting possessive determiners as tokens (rather than types) allows us to identify elaborate body language descriptions where the same word is repeated, as in "turning up ***his*** spectacles on ***his*** forehead and rubbing ***his*** hands" (setting A). In contrast, lines containing differently gendered determiners typically refer to intimate actions between characters, as in "putting ***her*** fair arm round ***his*** neck" (setting D).

The results of our case study, which will be presented in full detail at the conference, allow us to assess the practical usefulness of different types of concordance operations. The FlexiConc library provides a transparent account of the underlying algorithms, and hence the choices that analysts make to highlight different aspects of patterns. Thus, we can critically think about the process involved in 'reading' concordances. By using German and English, we additionally reflect on how concordance reading choices vary across languages.

References

Finlayson, Natalie, Stephanie Evert, Michaela Mahlberg, and Alexander Piperski (2024), Deutsche Romane des 19. Jahrhunderts (DE19): A nineteenth-century reference corpus of German novels for contrastive analysis, *Reading Concordances in the 21st Century Blog*, University of Birmingham. Available at: https://blog.bham.ac.uk/rc21/2024/04/26/deutsche-romane-des-19-jahrhunderts-de19-a-nineteenth-century-reference-corpus-of-german-novels-for-contrastive-analysis/

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel (2014), The Sketch Engine: Ten years on, *Lexicography* 1(1), 7–36. Available at: http://the.sketchengine.eu

Mahlberg, Michaela, Peter Stockwell, Viola Wiegand, and Jamie Lentin (2020), CLiC 2.1: Corpus Linguistics in Context. Available at: https://clic.bham.ac.uk

O'Donnell, Matthew Brook (2008), KWICGrouper – Designing a tool for corpus-driven concordance analysis, *International Journal of English Studies* 8(1), 107–122. Available at: https://revistas.um.es/ijes/article/view/49121

Coding situations involving non-canonical posture in present-day English

Thomas Egan
(University of Inland Norway)

This presentation contrasts two constructions, both of which contain two verbs. One of the verbs is a posture, or stance/positional, verb (hereafter PV), the other a verb coding an activity or state. In both constructions at least one of the two verbs is a gerund-participle. An example of one construction is *stood staring*, an example of the other is *stared standing*. In the first construction the PV is often said to have been subject to a degree of discursive backgrounding with concomitant semantic bleaching. As Quirk *et al.* put it, the PVs *lie*, *sit* and *stand* 'can take […] an *-ing* clause , with consequent weakening of the primary meaning of the main verb' (1985: 506. See also Huddleston and Pullum 2002: 1224, Newman and Rice 2004: 374, Ebeling 2015: 39 and Fraser 2018: [25]).

The hypothesis investigated in this presentation is grounded in the assumption that the PV in the 'PV + V-*ing*' construction is discursively secondary to the second verb, which bears the primary discourse focus (Boye and Harder 2012: 7-8). This backgrounding has led to the PV being bleached to such an extent that it is no longer suitable for encoding situations in which the speaker wishes to emphasise the posture of the referent of the syntactic subject. Thus the non-canonical posture of standing for sleeping, in the case of humans and many animals, warrants greater discursive emphasis than is conferred by first position in the

'PV + V-*ing*' construction. The requisite focus may, however, be conveyed by changing the order of the two verbs. The vertical sleeper does not normally *stand sleeping* in English. Rather they *sleep standing (up)*. There are over 30 examples of the latter with a human subject in COCA, compared to none of *stand sleeping*. The question arises as to whether this sort of distribution is characteristic of situations involving unusual posture. This leads to the following research question:

> Does the suitability of the subject's posture for the activity in question influence the distribution of the 'PV + V-*ing*' and 'V + PV-*ing*' constructions in English?

To explore this question, I present data from COCA for all three cardinal PVs with a selection of states and activities, the default postures for which are lying, sitting or standing (Newman and Rice 2004: 352, 368-370). I searched COCA for all examples of the three PVs followed by an *-ing* form of the V2 within two places. The reason for allowing for a space between the two verbs is to accommodate very common adverbs such as 'here', 'there', 'down' and 'up'. I then downloaded all examples of the second (now first) verb followed by an *-ing* form of the PV within two places. I compared the two sets of concordances in an effort to ascertain whether there is indeed a correlation between the suitability of the posture assumed and the construction chosen to encode the situation. In this presentation I present the results of this investigation.

References

Boye, Kasper, and Peter Harder (2012), A usage-based theory of grammatical status and grammaticalization, *Language* 88(1), 1–44.

Ebeling, Signe Oksefjell (2015), A contrastive study of Norwegian pseudo-coordination and two English posture-verb constructions, in S. O. Ebeling, and H. Hasselgård (eds), *Cross-Linguistic Perspectives on Verb Constructions*, Newcastle upon Tyne: Cambridge Scholars Publishing, 29–57.

Fraser, Katie (2018), Polysemous posture in English: A case study of non-literal meaning, in A. A. Spalek, and M. Gotham (eds), *Approaches to Coercion and Polysemy*, *Oslo Studies in Language* 10(2), 9–28.

Huddleston, Rodney D., and Geoffrey K. Pullum (2002), *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press.

Newman, John, and Sally Rice (2004), Patterns of usage for English sit, stand and lie: A cognitively inspired exploration in corpus linguistics. *Cognitive Linguistics* 15(3), 351–396.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985), *A Comprehensive Grammar of the English Language*, London: Longman.

Bootstrapping keywords and collocations in R

Stephanie Evert
(FAU Erlangen-Nürnberg)

Keyword and collocation analysis are two fundamental techniques of corpus linguistics with useful applications in many areas, including computational lexicography, studies of linguistic variation, phraseology, language learning, and computer-assisted discourse studies. Keywords refer to lexical items that are significantly more frequent in a given target corpus than in a reference corpus; collocations are words that co-occur with significantly higher frequency than expected by chance. Such patterns can be quantified by various statistical keyness or association measures, which are then used for ranking keyword or collocation candidates. Some measures aim to adjust for uncertainty – due to the fact that corpora are samples of language – using statistical hypothesis testing. One example is the recently suggested LRC measure (Evert 2022), which even corrects for multiple testing (cf. Gries 2024: 273f).

Both keywords and collocations are strongly affected by the fact that corpora aren't random samples of tokens (as most hypothesis tests assume) but rather samples consisting of entire texts or text fragments (e.g. Evert 2006; Gries 2024: 274). The known tendency of words (and co-occurrences of words) to repeat within the same text, also known as *term clustering* (Church 2000), leads to inflated estimates for the significance of frequency differences because the resulting sampling distribution has much greater

variability than the binomial distribution for token-level samples (cf. the striking visualisation of Lijffijt et al. 2016: 377, Fig. 1). This has a particularly profound effect on low-frequency data such as word co-occurrences.

Recent methodological work in corpus linguistics strongly recommends the use of bootstrapping approaches (Efron & Tibshirani 1993) that estimate the sampling distribution empirically by resampling corpora with replacement at the level of entire texts (Lijffijt et al. 2016; Gries 2022). However, such techniques are rarely applied in practice, which can be traced to several factors:

1. none of the standard corpus software tools implement bootstrapped keyness and collocation analysis, so researchers are left to their own devices and programming skills;
2. expositions of the methodology in research papers are often very technical, making it difficult for readers to understand exactly what data and algorithms are required;
3. the appropriate interpretation of bootstrapped sampling distributions remains unclear, with research papers often focused on visualising these distributions rather than explaining how they can be integrated into keyness and collocation scores; and
4. bootstrapping algorithms are computationally expensive, especially when applied to large data sets of keyword or collocation candidates, so substantial programming experience is needed to carry them out with sufficent efficiency (cf. Gries 2024: 275-297).

The purpose of this contribution is to provide a practical guide to bootstrapped keyword and collocation analysis in R, covering all the challenges addressed above: (i) how to obtain suitable frequency data from standard corpus software tools; (ii) how to perform bootstrapping efficiently with relatively simple R code; and (iii) how to interpret the bootstrapping results and integrate them sensibly into keyword and collocation rankings. Techniques will be illustrated on several example data sets, discussing the differences between traditional and bootstrapped results. Thoroughly documented and reproducible code examples will be made available as an online supplement, with some ready-made functions integrated into an open-source R package.

References

Efron, Bradley, and Robert Tibshirani (1993), *An Introduction to the Bootstrap*, *Monographs on Statistics & Applied Probability* 57, Boca Raton: Chapman & Hall/CRC.

Evert, Stefan (2006), How random is a corpus? The library metaphor, *Zeitschrift für Anglistik und Amerikanistik* 54(2), 177–190.

Evert, Stefan (2022), Measuring keyness, in *Digital Humanities 2022: Conference Abstracts*, Tokyo (online), 202–205. Available at: https://osf.io/cy6mw/

Gries, Stefan Th. (2022), Toward more careful corpus statistics: Uncertainty estimates for frequencies, dispersions, association measures, and more, *Research Methods in Applied Linguistics* 1(1), 100002.

Gries, Stefan Th. (2024), *Frequency, Dispersion, Association, and Keyness*, *Studies in Corpus Linguistics* 115, Amsterdam and Philadelphia: John Benjamins.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2016), Significance testing of word frequencies in corpora, *Digital Scholarship in the Humanities* 31(2), 374–397.

Indirect interrogatives and embedded inversion in Early Modern Irish (and British) English: Superstratum retention, substratum transfer, construction types

Teresa Fanego
(University of Santiago de Compostela)

This presentation examines *embedded inversion* (EI), a label applied to the inversion of verb and subject in indirect interrogative clauses:

1. You may ask **why did Pope Damasus feel** that this revision was necessary. (ICE-Ireland; Kallen 2013:77)

2. Sir vallentine asked him […] **in what maner would he Imploy them**

(1641 Depositions; Examination of James Linch FitzSteephen; Galway, 20/1/1652; MS 830/fol. 164v)

Indirect interrogatives are clauses in complement function which are licensed by an appropriate head (*ask*, *know*, etc.). They fall into two main categories, WH- interrogatives, as above, which are introduced by a WH-word, and YES/NO interrogatives, which in Standard British English (BrE) are introduced by *if* or *whether*.

EI is a frequent feature of Irish English (IrE) (Corrigan 2010:69-70; Kallen 2013:77-80), which differs significantly in this regard from BrE, where dependent interrogatives usually employ non-inverted order:

(3) I don't know **which she prefers**.

Although constructions with EI are reasonably common among varieties of English (eWAVE 3.0, feature 227), the high rate of attestation of EI in IrE has long been interpreted as due to substratal influence from Irish (Filppula 1999:169-179; Hickey 2007:276). Irish Gaelic, the Celtic language spoken in Ireland, retains the word order of direct questions also in dependent interrogatives.

The few studies (Davydova et al. 2011:306-316; Amador-Moreno 2019:168-199) addressing EI from a historical perspective have focused only on the Late Modern period, because the textual record of IrE prior to the 18th century is very scant; they date the earliest occurrences of EI to 1820. However, the availability since 2010 of a digital edition of the 1641 Depositions has opened new possibilities for research on EI in Early Modern IrE.

The 1641 Depositions are a compilation of witness testimonies recorded after the 1641 rebellion in Ireland, an uprising initiated in October 1641 by Catholic gentry and military officers. The testimonies were collected by a state-appointed commission, and document the alleged crimes committed by the Catholic Irish insurgents. This presentation proposes to investigate a set of 689 dependent interrogatives retrieved from a 800,617 word corpus compiled from the 1641 Depositions. A comparative analysis of 503 dependent interrogatives in BrE will also be carried out, this using the witness depositions' component (172,940 words) contained in the *Corpus of English Dialogues 1560-1760* (CED; Kÿto & Walker 2006). Our specific aims are as following:

i. To show that the 1641 Depositions, despite the bias of their compilation, constitute a valuable source of evidence on 17th-century IrE. To date, they have hardly been subject to linguistic analysis.

ii. To offer an account of the features of EI in the corpora chosen for study, and examine its interrelation with construction type (WH- versus YES/NO-).

iii. To assess how the corpus results relate to the two main explanatory factors proposed in the literature for the use of EI in IrE, namely (a) influence of the superstrate, given the existence in historical BrE of some degree of use of EI; and (b) influence of the substrate language (Irish Gaelic).

References

1641 Depositions, Trinity College Dublin. https://1641.tcd.ie/

Amador-Moreno, Carolina P. (2019), *Orality in Written Texts. Using Historical Corpora to Investigate Irish English 1700-1900,* London & New York: Routledge.

Corrigan, Karen P. (2010), *Irish English, Volume 1 – Northern Ireland,* Edinburgh: Edinburgh University Press.

Davydova, Julia, Michaela Hilbert, Lukas Pietsch, and Peter Siemund (2011), Comparing varieties of English: Problems and perspectives, in Peter Siemund (ed), *Linguistic Universals and Language Variation*, Berlin: Mouton de Gruyter, 291–323.

Filppula, Markku (1999), *The Grammar of Irish English: Language in Hibernian Style,* London & New York: Routledge.

Hickey, Raymond (2007), *Irish English: History and Present-day Forms,* Cambridge: Cambridge University Press.

Kallen, Jeffrey L. (2013), *Irish English Volume 2: The Republic of Ireland,* Berlin: Mouton de Gruyter.

Kortmann, Bernd, Kerstin Lunkenheimer, and Katharina Ehret (2020), *The Electronic World Atlas of Varieties of English*. Available online at http://ewave-atlas.org; accessed on 2024-12-09.

Kytö, Merja and Terry Walker (2006), *Guide to a Corpus of English Dialogues 1560–1760*. Acta Universitatis Upsaliensis 130, Uppsala: Uppsala Universitet.

Identifying a structural linguistic epicentre: An analysis of adjective comparison in South Asian Englishes

Nina Funke
(Justus Liebig University Giessen)

Following Hoffmann et al. (2011) and Hundt (2013), a linguistic epicentre needs to meet two criteria: endonormative stabilization, as defined by Schneider (2007), and "the potential to serve as a model of English for (neighbouring?) countries" (Hundt 2013: 185). The present study adapts a stance taken by Schneider (2022) claiming that "epicentral influence is to be seen not as an all-or-nothing effect but as a 'prototypical' concept, a relationship which may hold to a lesser and stronger extent and which in turn is composed of and can be detected by a range of composite factors" (Schneider 2022: 472). In taking this stance, this study seeks to answer the following research question:

Which among the South Asian varieties of English is the most prototypical epicentre with regard to the adjective comparison alternation?

To answer this question, this study draws a total of 3536 analytically and synthetically compared adjectives from the SAVE (Bernaisch et al. 2011), SAVE2020 (Bernaisch et al. 2021), COCA (Davies 2008-) and NOW (Davies 2016-) corpora. These corpora include data from American, and British English as well as the six South Asian varieties Bangladeshi, Indian, Maldivian, Nepali, Pakistani, and Sri Lankan English. The study identifies a set of structural factors that determine epicentral influence over two points in time (i.e., 2000s and 2020 and how they can be operationalized with the help of random forests using the example of the adjective comparison alternation. Furthermore, this study shows how two groups of structural factors, i.e., those operationalizing endonormativity on one hand, e.g., the worst predictions by American/British English for the potential epicentre in both time periods, and the potential to serve as a model on the other hand, e.g., the most improved predictions on average for the varieties in the region from the earlier to the later time period, may be used to make a statement about the prototypicality of a potential epicentre of South Asian Englishes compared to other varieties under investigation. In this vein, this study identifies Indian English as having the strongest influence over the other South Asian varieties and Nepali English being the least influenced by American and British English when it comes to the comparison of adjectives. Overall, Indian English seems to be the most prototypical epicentre of the adjective comparison alternation among the South Asian varieties. However, as Götz (2022: 357) argues, "not all types of linguistic features seem to be prone to a successful spread through epicentres." Therefore, the present results may benefit from analyses of additional linguistic features.

References

Bernaisch, Tobias, Benedikt Heller, and Joybrato Mukherjee (2021), *Manual for the 2020-Update of the South Asian Varieties of English (SAVE2020) Corpus*, Version 1.1, Giessen: Justus Liebig University, Department of English.

Bernaisch, Tobias, Christopher Koch, Joybrato Mukherjee, and Marco Schilk (2011), *Manual for the South Asian Varieties of English (SAVE) Corpus: Compilation, Cleanup Process, and Details on the Individual Components*, Giessen: Justus Liebig University, Department of English.

Davies, Mark (2008–), *The Corpus of Contemporary American English (COCA)*. Available at: https://www.english-corpora.org/coca/

Davies, Mark (2016–), *Corpus of News on the Web (NOW)*. Available at: https://www.english-corpora.org/now/

Götz, Sandra (2022), 'Epicentral Influences of Indian English on Nepali English', *World Englishes*, 41: 347–360. https://doi.org/10.1111/weng.12582

Hoffmann, Sebastian, Marianne Hundt, and Joybrato Mukherjee (2011), 'Indian English – An Emerging Epicentre? A Pilot Study on Light Verbs in Web-Derived Corpora of South Asian Englishes', *Anglia*, 129: 258–280. https://doi.org/10.1515/angl.2011.083

Hundt, Marianne (2013), 'The Diversification of English: Old, New and Emerging Epicentres', in Daniel Schreier and Marianne Hundt (eds), *English as a Contact Language*, Cambridge: Cambridge University Press, 182–203. https://doi.org/10.1017/CBO9780511740060.011

Schneider, Edgar W. (2007), *Postcolonial English: Varieties Around the World*, Cambridge: Cambridge University Press.

Schneider, Edgar W. (2022), 'Parameters of Epicentral Status', *World Englishes*, 41: 462–474. https://doi.org/10.1111/weng.12589

Exploring certainty and possibility: Epistemic modality in English and Italian popular science books

Viviana Gaballo and Sara Gesuato
(Padua University)

This study explores the use of epistemic modality—linguistic expressions of certainty, possibility, and predictability—in popular science books written in English and Italian, with the aim of understanding how scientific knowledge is communicated to lay audiences across languages and cultures. Drawing on Halliday's systemic-functional framework (1970, 2004) and insights from cross-linguistic and genre-specific studies (e.g., Biber & Finegan 1988; Carrió Pastor 2012; Rozumko 2017), the research investigates the frequency and distribution of modal expressions that signal varying degrees of speaker/writer commitment to truth in a comparable corpus of approximately one million words in each language. The sub-corpora encompass texts in the fields of physics, evolutionary biology, astronomy, linguistics, and cultural history, though they are not identically balanced across disciplines.

The study focuses on a set of prototypical modal auxiliaries, adverbs, and adjectives that signal epistemic stance in both languages (e.g., *may, might, probably* in English; *potrebbe, forse, sicuramente* in Italian), identified based on prior literature (Palmer 2001; Colella 2015; Giannakidou & Mari 2018; La Forgia 2018). The goal was to determine whether semantically comparable resources are used with similar frequency and in comparable syntactic-semantic contexts across the two corpora.

Findings indicate that epistemic modality is generally infrequent in both English and Italian popular science texts. However, notable cross-linguistic differences emerge in the preferred means of expressing epistemic stance. English texts more frequently employ modal auxiliaries such as *may*, *might*, and *could*, which align with conventional patterns in English-language academic and scientific discourse. Italian texts, in contrast, more often utilize conditional verb forms such as *potrebbe* ('could/might') or periphrastic constructions like *sembra che* ('it seems that') and *è possibile che* ('it is possible that'). These findings suggest that while English tends to lexicalize epistemic judgments through auxiliary verbs, Italian often relies on syntactic strategies and mood- inflected verb forms.

The results underscore that popular science writing tends to balance epistemic caution with communicative clarity, aiming to convey authority while remaining accessible. In this respect, both corpora show a preference for relatively explicit and subjective epistemic markers (Nyuts 2001), supporting previous claims that scientific communication for general audiences leans toward assertiveness, even when hedging is present (Vold 2006; Díaz Rojo 2007; Kranich 2009; Hacquard & Wellwood 2012; Yang et al. 2015; Liu 2024). Furthermore, the study suggests that language- specific grammatical resources and cultural conventions shape how scientific uncertainty is articulated, with Italian perhaps favoring implicit or morphologically encoded modality more than English.

Overall, this contrastive analysis contributes to a deeper understanding of how epistemic modality operates in cross-linguistic science communication. It offers implications for the translation of scientific texts, for science education, and for improving intercultural accessibility in public-facing science discourse.

The findings also pave the way for future research on modality in other language pairs and genres, particularly in the context of increasing globalization of science communication.

References

Biber, Douglas, and Edward Finegan (1988), Adverbial stance types in English, *Discourse Processes* 11(1), 1–34.

Carrió Pastor, María Luisa (2012), A contrastive analysis of epistemic modality in scientific English, *Revista de Lenguas para Fines Específicos* 18, 115–132.

Colella, Gianluca (2015), Marcatori epistemici avverbiali in italiano contemporaneo, *La Lingua Italiana: Storia, Strutture, Testi* XI, 137–162.

Díaz Rojo, José Antonio (2007), La modalidad epistémica asertiva como recurso retórico en noticias científicas: el caso del hallazgo paleontológico del Hombre de Orce, *Revista de Lingüística y Lenguas Aplicadas* 2, 17–27.

Giannakidou, Anastasia, and Alda Mari (2018), The semantic roots of positive polarity: Epistemic modal verbs and adverbs in English, Greek and Italian, *Linguistics and Philosophy* 41, 623–664.

Hacquard, Valentine, and Alexis Wellwood (2012), Embedding epistemic modals in English: A corpus-based study, *Semantics and Pragmatics* 5(4), 1–29.

Halliday, Michael A. K. (1970), Functional diversity in language as seen from a consideration of modality and mood in English, *Foundations of Language* 6(3), 322–361.

Halliday, Michael A. K. (2004), *An Introduction to Functional Grammar*, London: Arnold.

Kranich, Svenja (2009), Epistemic modality in English popular scientific texts and their German translations, *Trans-kom* 2(1), 26–41.

La Forgia, Francesca (2021), Epistemic, evaluative, speech act adverbs and Italian political language, *Linguistik Online* 92(5), 145–172.

Liu, Tianqi (2024), A comparative study on epistemic modality in linguistic research article conclusions, *Theory and Practice in Language Studies* 14(2), 476–484.

Nuyts, Jan (2001), Subjectivity as an evidential dimension in epistemic modal expressions, *Journal of Pragmatics* 33, 383–400.

Palmer, Frank R. (2001), *Mood and Modality*, Cambridge: Cambridge University Press.

Rozumko, Agata (2017), Adverbial markers of epistemic modality across disciplinary discourses: A contrastive study of research articles in six academic disciplines, *Studia Anglica Posnaniensia* 52(1), 73–101.

Yang, An, Shu-Yuan Zheng, and Guang-Chun Ge (2015), Epistemic modality in English-medium medical research articles: A systemic functional perspective, *English for Specific Purposes* 38, 1–10.

Vold, Eva Thue (2006), Epistemic modality markers in research articles: A cross-linguistic and cross-disciplinary study, *International Journal of Applied Linguistics* 16(1), 61–87.

Non-decomposable affixed words produced by L1 Spanish EFL writers

Roger Gee, Kathleen Jogan and Mary Karen Jogan
(Holy Family University, University of Arkansas, Albright College)

Over 30 years ago, Bauer and Nation (1993) proposed a teaching framework for affixes that is still referenced today. More recently, Nation and Bauer (2023) described how their framework can be used to teach morphological awareness. Importantly, they cautioned that affixed words that are more frequent than their stems are treated as wholes by native speakers rather than decomposable words. That is, they are non-decomposable. Based on the work of Hay (2001), Nation and Bauer argued that it is the relative frequencies of the stem and the affixed words that is more important than the simple frequency of the affixed words. Hay had asked participants in her study to intuitively identify which one of a word pair was easier to decompose. Affixed words less frequent than their bases were consistently chosen as easier to

decompose, supporting the relative frequency argument. Recently, Ur (2022) relied on only her intuition to identify words that were non-decomposable, "whose meaning could not be inferred from a combination of the meanings of the affix and that of the baseword …. Examples are *inter + view*, *hard + ly*, *re + place*" (p. 332).

Hay also used semantic drift, when there is a meaning change over time, as a criterion for decomposability. When a dictionary definition of the affixed word contains the root, there was no semantic drift and the word was decomposable, as in *seater/with enough seats for the particular number of people.* Words whose root was not given in a definition were considered to have experienced semantic drift and were non-decomposable, as in *barely/by the smallest amount.*

This presentation will address two questions about a set of words intuitively chosen by three experienced L1 English teachers as being difficult to decompose:

1. Do the words identified as being non-decomposable have a relative frequency between the root and the derived word that is greater than one?
2. Do the words identified as non-decomposable show semantic drift as determined by dictionary definitions?

The data was drawn from affixed words used by Spanish L1 EFL learners in COREFL (Lozano et al., 2020), a corpus of 436 texts containing 85,856 words. Their proficiencies ranged from A1-C2. Affixed words were identified with Morpholex (Cobb, 2022) using Bauer and Nation's (1993) levels, but the data was aggregated without regard to affix levels. Non-decomposable affixed words were intuitively identified by three experienced L1 English teachers.

Relative frequencies of the root and affixed words were determined using frequency data from COCA (Davies, 2008-). Semantic drift was determined using definitions of the affixed words in the Cambridge Dictionary (https://dictionary.cambridge.org/us/).

A binomial test with a Bonferroni correction showed that the proportion of relative frequency scores greater than 1 was significantly less than expected (34.8%, p = .210). The proportion of words experiencing semantic drift was significantly greater than expected (87.0%, p = <.001). The results suggest that intuition by teachers may be superior to relative frequency as a measure of how affixed words are processed and less time consuming than using dictionary definitions.

References

Bauer, Laurie and Paul Nation (1993), Word families, *International Journal of Lexicography*, 6(4), 253–279.

Cambridge Dictionary, Cambridge University Press. https://dictionary.cambridge.org/us/

Cobb, Tom (2022), Morpholex. (https://www.lextutor.ca/)

Davies, M. (2008-). *The Corpus of Contemporary American English*. www.english-corpora.org/coca/.

Hay, Jennifer (2001), Lexical frequency in morphology: Is everything relative? *Linguistics*, 39(6), 1041-–1070.

Lozano, Cristóbal, Ana Díaz-Negrillo, and Marcus Callies (2020), Designing and compiling a learner corpus of written and spoken narratives: COREFL, in C. Bongartz and J. Torregrossa (eds), *What's in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy,* Peter Lang, 21–46. https://doi.org/10.3726/978-3-653-05182-7

Nation, Paul and Laurie Bauer (2023), What is morphological awareness and how can you develop it? *Language Teaching Research Quarterly,* 33, 80–98.

Ur, Penny (2022), How useful is it to teach affixes in intermediate classes? *ELT Journal*, 76(3), 330–337.

The eye of the needle, the needle's eye or the needle eye? Shared and unshared semantic preferences in English complex noun phrases

Sara Gesuato
(Padua University)

English has three primary ways to indicate a semantic connection between two NPs: the *of*-genitive, the *'s*-genitive and compounding. These structures can express various sense relationships, including possessor-possessee, agent-action, experiencer-experience, patient-action, beneficiary-benefactor, part-whole, property-entity and others.

The choice between encoding options is influenced by structural, morpho-phonological, semantic and pragmatic considerations, including the need for information packaging (Almazan Ruiz 2021), the final sound (Szmrecsanyi, Hinrichs 2008) and the animacy (Almazan Ruiz 2021; Szmrecsanyi et al. 2016) of the dependent noun. The notion encoded by the head (Breban et al. 2019; Bruton 2002) and its definiteness (Bruton 2002) also play a role. The 's-genitive commonly encodes agent-action, experiencer-experience (Almazan Ruiz 2021) and possessor-possessee relationships (Breban et al. 2019). Compounding often represents the name-entity and location-entity relationships (Breban et al. 2016), while the *of*-genitive typically occurs with non-human dependents (Bruton 2002).

This study used corpus data and elicited data to examine eight variants of structures consisting of a nominal head and a nominal modifier, both definite and indefinite, variously combined: *a Noun Noun; the Noun Noun; a Noun's Noun; the Noun's Noun; a(n) N of a(n) N; the Noun of the Noun; a(n) Noun of the Noun;* and *the Noun of a Noun*. Corpus data were meant to reveal the frequencies of use of, and the range of semantic relationships encoded in, the above variants in colloquial discourse. Elicited data were used to explore the compatibility of a given semantic relationship with more than one variant.

First, 400 examples of each variant were collected from the *Film Corpus* (21 million words), available through *SketchEngine*, which exemplifies naturalistic, informal scripted conversation and short descriptive-narrative text segments. From each set, 5 instances were selected in their sentential contexts, which exemplified singular countable head nouns and dependents; then, the original NPs were replaced with blank slots. Each blank was paired with eight versions of NPs representing the variants listed above. Then, 30 native speakers recruited via *Prolific* rated the naturalness of these variants in an online questionnaire.

Next, the semantic relationship encoded in the corpus examples were identified, following Breban et al. (2019). Four findings emerged: A) the four most common variants were *the Noun's Noun, the Noun Noun, A Noun Noun,* and *the Noun of the Noun*, in this order; B) the *'s*-genitive construction preferred a human agentive dependent, whether definite or indefinite, primarily encoding possessor-possessee or, less frequently, agent-action relationships (e.g. "Aldo's head; Harvey's handshake; a child's toy; a doctor's care"); C) the *of*-genitive construction often encoded: i) part-whole relationships with inanimate entities, when both the head and the dependent were definite NPs (e.g. "the edge of the building; the rest of the team"); ii) part-whole, agent-action or property-entity relationships, when only the head was definite (e.g. "the mouth of a whale; the call of a meadow lark; the grace of a Samurai"); or iii) part-whole and property-entity relationships, when both NPs were indefinite (e.g. "a fraction of an inch; a dream of a girl"); and D) compound NPs expressed property/component-entity or the beneficiary/goal-entity relationships involving inanimate entities, with definite or indefinite NPs (e.g. "a news station; a chocolate bar; a patio table; the storage unit; the mission objective; the cactus garden").

The elicited data revealed preferences similar to those found in the corpus data, while also showing raters' acceptance of alternative wordings: thus, I) in sentences originally featuring *of*-genitive variants, *of*-genitives were rated more natural than compounds and especially *'s*-genitives, and they were also compatible with both definite and indefinite nouns; II) in sentences with original *'s*-genitive variants, *'s*-genitives and *of*-genitives were rated as equally natural when the dependent was definite, while compounds were generally rated as unnatural; and finally, III) in sentences with original compound variants, compounds were highly, but not consistently, preferred, and were occasionally replaceable by *of*-genitives.

Consistent with Breban et al. (2019), morphological-semantic associations in the above nominal structural variants appeared to be both prototypical and underspecified, with their interpretation relying on contextual inference.

References

Almazán Ruiz, Encarnación (2021), 'El Genitivo Agente como la Interpretación Semántica más Prototípica en Inglés: Un Estudio Corpus', *Lingüística y Literatura* 79, 112–131.

Breban, Tine, Julia Kolkmann, and John Payne (2019), 'The Impact of Semantic Relations on Grammatical Alternation: An Experimental Study of Proper Name Modifiers and Determiner Genitives', *English Language and Linguistics* 23(4), 797-826. https://doi.org/10.1017/S1360674319000234

Bruton, Anthony Stewart (2002), 'Recoding and Reorganizing Grammatical Form by Meaning: The English Genitive as an Example', *System* 30, 237–250.

Szmrecsanyi, Benedikt, and Lars Hinrichs (2008), 'Probabilistic Determinants of Genitive Variation in Spoken and Written English: A Multivariate Comparison Across Time, Space, and Genres', in Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta, and Minna Korhonen (eds), *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*, Amsterdam/Philadelphia: John Benjamins, 291–309.

Szmrecsanyi, Benedikt, Douglas Biber, Jesse Egbert, and Karlien Franco (2016), 'Toward More Accountability: Modeling Ternary Genitive Variation in Late Modern English', *Language Variation and Change* 28(1), 1–29.

The evolving language of blogs: A diachronic corpus-based analysis

Marianna Gracheva, Daniel Keller and Jesse Egbert
(Friedrich-Alexander-University, Western Kentucky University, Northern Arizona University)

In the text linguistic tradition, registers are culturally recognized varieties of texts, associated with the communicative situation of use (emails, textbooks, conversations, etc.). The situation of use in turn calls for particular linguistic characteristics. Linguistic features frequent in a register are thus not arbitrary but are a direct response to the situation of this register and perform communicative functions necessitated by the situation (Biber & Conrad, 2019). At the same time, the texts of any given register are not homogenous in their linguistic or situational characteristics (e.g., Biber et al., 2020; Egbert & Gracheva, 2023; Wood, 2023; Goulart, 2024): As the situations in which texts of a register are created vary, so do the linguistic features that are frequent in those texts. (Biber & Egbert, 2023).

These functional links between situation and language raise new questions about registers' evolution as cultural constructs. First, as communicative situations evolve, language must reflect language users' adaptations to the new situations of use. Second, varying degrees of register-internal variability at different points of registers' existence could reflect language users' degrees of convergence (or lack thereof) on certain communicative and linguistic register norms.

This study begins investigating these questions focusing on blogs—a relatively new register but characterized by a rapidly evolving technological landscape (Miller & Shepherd, 2004, 2009) and whose full life cycle is available for scientific study—from its inception in the late 90s to the present day (corpus under analysis: N_{texts} = 2,452; N_{tokens} = ~4,000,000; source: Blogspot.com; years: 1999–2023, approx. 100 texts/year). We address the following research questions (RQ):

1. What linguistic features of blogs have become more or less frequent over time?

2. How does the linguistic stability of blogs—the degree of variation in the linguistic characteristics—change over time?

To address RQ 1, we compute rates of occurrence for a wide range of lexico-grammatical features in each text (the corpus is tagged for 150 linguistic features with an updated version of the Biber tagger; Biber, 1988) and employ corresponding feature analysis (Egbert, 2024) to examine correlations between rates of

occurrence of linguistic features and text dates. The analysis revealed that features associated with oral/interactive communication and clausal elaboration (e.g., pronominal references, various verb types) have increased over time, while features of information density (e.g., nouns) and abstractedness (e.g., passive constructions) have decreased. Blogs have also become increasingly present-oriented, and features of narrativity (e.g., past tense) have shown a downward trend.

In response to RQ 2, we present analyses of coefficients of variation (Segalowitz & Segalowitz, 1993) as a way of examining linguistic stability and show that variation in the majority of blog features has decreased, pointing to an increased stability in the linguistic norms of the register. We situate the study within the theme of the conference—exploring the past and mapping the future—by discussing our findings as language users' reconceptualization of the register, adapting to its communicative affordances, and developing a corresponding linguistic profile.

References

Biber, Douglas (1988), *Variation across Speech and Writing*, Cambridge: Cambridge University Press.

Biber, Douglas, and Susan Conrad (2019), *Register, Genre, and Style*, Cambridge: Cambridge University Press.

Biber, Douglas, and Jesse Egbert (2023), What is register? *Register Studies* 5(1), 1–22.

Biber, Douglas, Jesse Egbert, and Daniel Keller (2020), Reconceptualizing register in a continuous situational space, *Corpus Linguistics and Linguistic Theory* 16(3), 581–616.

Egbert, Jesse (2024, September 14), The text-linguistic (r)evolution [Plenary talk], *American Association for Corpus Linguistics*, Eugene, OR, USA.

Egbert, Jesse, and Marianna Gracheva (2023), Linguistic variation within registers: Granularity in textual units and situational parameters, *Corpus Linguistics and Linguistic Theory* 19(1), 115–143.

Goulart, Larissa (2024), *Variation in University Student Writing: A Communicative Text Type Approach*, Amsterdam: John Benjamins Publishing Company.

Miller, Carolyn R. and Dawn Shepherd (2004), Blogging as social action: A genre analysis of the weblog, in L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff and J. Reymann (eds), *Into the Blogosphere: Rhetoric, Community, and the Culture of Weblogs*, Minneapolis: University of Minnesota.

Miller, Carolyn R. and Dawn Shepherd (2009), Questions for genre theory from the blogosphere, in J. Giltrow and D. Stein (eds), *Genres in the Internet: Issues in the Theory of Genre*, Amsterdam: John Benjamins Publishing Company, 263–290.

Segalowitz, Norman S., and Sidney J. Segalowitz (1993), Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition, *Applied Psycholinguistics* 14(3), 369–385.

Wood, Margaret (2023), *Communicative Function and Linguistic Variation in State Statutory Law* [Doctoral dissertation, Northern Arizona University].

Are corpus annotators culturally biased?

Wenwen Guan
(University of Amsterdam)

"*I think*" serves as an epistemic modality marker, recognized for its functional flexibility. Numerous studies have studied its functions across various contexts, with the majority employing corpus-based methodologies (e.g., Aijmer, 1997; Li & Pang, 2022; Simon-Vandenbergen, 2000; Wang, 2020; Zhang, 2014). Even when provided with certain contextual cues summarized in aforementioned studies, a systematic disagreement emerged among my annotators as they annotated the functions of "*I think*" in the International Corpus Network of Asian Learners of English (ICNALE). This observation prompted me to further explore the paradoxical roles of "*I think*" in expressing certainty and uncertainty with an experimental approach.

To clarify the disagreement, the annotators held different views on the functions of "*I think*" in the same context: some identified it as a softening device, while others considered it to be the opposite. Similarly, I hesitate to endorse the finding in a previous study (Kobayashi, 2016) of the same corpus, which claimed that Japanese learners of English frequently used "*I think*" as a softening device. Considering that the ICNALE corpus collected participants' opinions on certain topics, I contend that "*I think*" in this opinion-sharing register appears to signal authoritative deliberation, as also argued by Simon-Vandenbergen (2000) in her analysis of political interviews. Given the established influence of cultural norms on the use of epistemic modality markers (Yu, 2019; Warchał, 2015), I propose a bold hypothesis: could the disagreement be attributed to the annotators' differing cultural backgrounds?

To address the question, a sociolinguistic experiment is designed in this study to investigate how individuals from varied cultural backgrounds evaluate the certainty conveyed by "*I think*". The experiment stimuli include the following linguistic factors to constitute the discourse contexts for "*I think*": co-occurring epistemic qualifiers (e.g., *might, possibly, should, definitely*), statement types (policy, evaluation, and fact), the position of "*I think*" in the sentence (initial, medial, and final), and its position in an argument (premise and conclusion). The participants consist of over 100 native and non-native speakers of English mainly from Europe and East Asia. The experiment aims to align as closely as possible with language use in real life, so no deliberate effort has been made to recruit professional annotators or linguistics students as participants. Comparing participants' responses within the same context will disclose whether "*I think*" is perceived differently.

This study endeavors to leverage an experimental approach to validate the causes of the inter-annotator disagreement, offering a novel supplementary method to mitigate annotation pitfalls. The findings will primarily contribute to developing annotation guidelines of "*I think*". More broadly, this experiment is expected to raise awareness of potential cultural biases among annotators when conducting corpus annotation, especially with regard to pragmatic terms. I have to admit the experiment has limitations. For example, the experiment is presented with written language as a simulation of genuine corpus annotation, although prosodic features can influence the hearers' perception. On second thought, however, this experiment could serve as a motivation for corpus linguists to consider annotation using multimodal data in certain cases.

References

Aijmer, Karin (1997), I think—An English modal particle, in T. Swan & O. J. Westvik (eds), *Modality in Germanic Languages*, De Gruyter Mouton, 1–48. https://doi.org/10.1515/9783110889932.1

Kobayashi, Yuichiro (2016), Investigating metadiscourse markers in Asian Englishes: A corpus-based approach, *Language in Focus* 2(1), 19–35. https://doi.org/10.1515/lifijsal-2016-0002

Li, Lanping, and Yang Pang (2022), A Corpus-Based Pragmatic Analysis of Discourse Marker *I Think*, *Open Access Library Journal* 9(10), 1–11. https://doi.org/10.4236/oalib.1109301

Simon-Vandenbergen, Anne-Marie (2000), The functions of *I think* in political discourse, *International Journal of Applied Linguistics* 10(1), 41–63. https://doi.org/10.1111/j.1473-4192.2000.tb00139.x

Wang, Qian (2020), A corpus-based contrastive analysis of *I think* in spoken Hong Kong English: Research from the International Corpus of English (ICE), *Australian Journal of Linguistics* 40(3), 319–345. https://doi.org/10.1080/07268602.2020.1823817

Warchał, Krystyna (2015), *Certainty and doubt in academic discourse: Epistemic modality markers in English and Polish linguistics articles*, Wydawnictwo Uniwersytetu Śląskiego.

Yu, Lee Seunghye (2019), *A cross-linguistic and cross-cultural study of stance markers in research articles in English and Korean.* Doctoral dissertation, University of Hawai'i at Manoa.

Zhang, Grace (2014), The elasticity of *I think*: Stretching its pragmatic functions. *Intercultural Pragmatics*, 11(2), 225–257. https://doi.org/10.1515/ip-2014-0010

Rhoticity in Bahamian English: A case of glocalization

Stephanie Hackert
(University of Munich)

A prominent feature of standard American English pronunciation is rhoticity, i.e., the occurrence of [r] in non-prevocalic positions, as in *four* and *fourth* (Labov 1972: 43-69), reflecting the historical presence of /r/ in all positions in English prior to the eighteenth century (cf., e.g., Hickey 2014). Rhotic pronunciations are stereotypically associated with American speech all over the world; they are highly salient, often explicitly commented on, and may call forth strong attitudinal reactions. With the exception of Barbados, the anglophone Caribbean is traditionally non-rhotic, but rhotic pronunciations appear to be on the rise (cf., e.g., Hackert 2004; Rosenfelder 2009; Deuber & Leung 2013; Kraus 2017; Irvine-Sobers 2018; Meer 2024), despite the fact that they are often seen as indicative of the threat posed by American lifestyles and culture to local forms of expression (Oenbring 2010: 52). All studies concur that, as elsewhere in the English-speaking world, rhoticity in the Caribbean is influenced by both language-internal and socio-stylistic factors such as preceding vowel (e.g., NURSE vs. START), word stress, gender, text type, etc.

The proposed paper investigates rhoticity in the spoken parts of ICE-Bahamas, the forthcoming Bahamian subcomponent of the International Corpus of English (ICE). It presents a multivariate analysis of the phenomenon in private face-to-face conversations as well as in broadcast discussions, interviews, news, and talks. Overall, rhotic pronunciations are particularly frequent among younger, highly educated female speakers in public formal situations in which a high level of correctness is aimed at, which suggests that realized /r/s are part of a new, local standard Bahamian accent. That said, even though they are superficially identical to a stereotypical feature of American English, it may be too simplistic to view rhotic pronunciations simply as a sign of cultural and linguistic Americanization. At least in the Bahamas, they might, in fact, be indicative of advanced decolonization and postcolonial nation building, which becomes clear when contextual cues, explicit comment, and attitudinal findings (Laube & Rothmund 2021) are taken into account. This ties in with findings from the lexical, orthographical, and grammatical levels of language (e.g., Deuber et al. 2022), which have shown that, despite their geographical proximity and long-standing demographic, cultural, institutional, and economic ties, the Bahamas are not undergoing wholesale linguistic Americanization. Rather, speakers are selectively taking up features of American English, molding them to fit the local linguistic ecology.

References

Deuber, Dagmar and Glenda-Alicia Leung (2013), Investigating attitudes towards an emerging standard of English: Evaluations of newscasters' accents in Trinidad, *Multilingua* 32(3), 289–319.

Deuber, Dagmar, Stephanie Hackert, Eva Canan Hänsel, Alexander Laube, Mahyar Hejrani, and Catherine Laliberté (2022), The norm orientation of English in the Caribbean: A comparative study of newspaper writing from ten countries, *American Speech* 97(3), 265–310.

Hackert, Stephanie (2004), *Urban Bahamian Creole: System and Variation,* Amsterdam: Benjamins.

Hickey, Raymond (2014), Vowels before /r/ in the history of English, in Simone E. Pfenninger, Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja, Marianne Hundt, and Daniel Schreier (eds), *Contact, Variation, and Change in the History of English,* Amsterdam: Benjamins, 95–110.

Irvine-Sobers, G. Alison (2018), *The Acrolect in Jamaica: The Architecture of Phonological Variation,* Berlin: Language Science Press.

Kraus, Janina (2017), *A Sociophonetic Study of the Urban Bahamian Creole Vowel System,* Ph.D. dissertation, Ludwig-Maximilians-Universität Munich, Available at https://edoc.ub.uni-muenchen.de/21689/1/Kraus_Janina.pdf (October 21, 2024).

Labov, William (1972), *Sociolinguistic Patterns,* Philadelphia: University of Pennsylvania Press.

Laube, Alexander and Janina Rothmund (2021), "Broken English", "Dialect" or "Bahamianese"? Language attitudes and identity in The Bahamas, *Journal of Pidgin and Creole Languages* 36(2), 362–394.

Meer, Philipp (2024), Variation and change in the NURSE vowel in Trinidadian English. An apparent-time analysis of adolescent and adult speakers, in Mirjam Schmalz, Manuela Vida-Manl, Sarah Buschfeld, and Thorsten Brato (eds), *Acquisition and variation in World Englishes: Bridging paradigms and rethinking approaches,* Berlin: de Gruyter, 279–306.

Oenbring, Raymond (2010), Corpus linguistic studies of standard Bahamian English: A comparative study of newspaper usage, *The International Journal of Bahamian Studies* 16, 51–62. DOI: 10.15362/ijbs.v16i0.124.

Rosenfelder, Ingrid (2009), Rhoticity in educated Jamaican English: An analysis of the spoken component of ICE-Jamaica, in Thomas Hoffmann and Lucia Siebers (eds), *World Englishes – problems, properties and prospects,* Amsterdam: Benjamins, 61–82.

A multifactorial analysis of the comparative alternation across World Englishes

Tjorven Halves
(University of Bonn)

The alternation between the analytic (*more silly*) and the synthetic (*sillier*) comparative is one of several extensively investigated grammatical alternation phenomena (e.g., Mondorf, 2009; Mondorf & Pérez-Guerra, 2016). Researchers have identified phonological, morphosyntactic, and usage-related constraints on the comparative alternation and suggested potential underlying causes, such as processing complexity and rhythmic alternation (e.g., Hilpert, 2008; Mondorf, 2009). However, these studies used monofactorial methods (e.g., Kytö & Romaine, 1997; Mondorf, 2009) or fixed-effects regression (Cheung & Zhang, 2016; Hilpert, 2008), disregarding two essential considerations: (i) phenomena investigated using corpuslinguistic data are never monofactorial (Gries, 2018) and (ii) the adjective lemma may have a significant confounding influence on the comparative alternation. The present study fills this research gap by analyzing the comparative alternation using mixed-effects regression, a multifactorial method that can account for the effect of the individual adjective.

Moreover, although interest in morphosyntactic variation in World Englishes has increased (e.g., Szmrecsanyi & Kortmann, 2009), the comparative alternation has not yet been investigated across many varieties of English. Researchers have examined grammatical alternations across varieties from a variationist perspective, inquiring whether their constraints vary across Englishes based on factors like L1 influence or nativization (e.g., Szmrecsanyi & Grafmiller, 2023). Studies found a significant overlap in the influence of the constraints across varieties, which indicates that these grammatical alternations may belong to a lexicogrammatical common core of English (e.g., Bernaisch et al., 2014; Heller et al., 2017). The present study contributes to this research by investigating the comparative alternation in 20 different Englishes. It answers the following research questions:

1. Which influence do phonological, morphosyntactic, and usage-related factors have on the choice between comparative variants?
2. To what extent does the influence of the phonological, morphosyntactic, and usage- related factors vary across World Englishes?

The analysis included 2,631,142 tokens from 110 different adjective lemmata from the GloWbE corpus (Davies & Fuchs, 2015). Eleven phonological, morphosyntactic, and usage- related constraints on the comparative alternation were examined, and mixed-effects regression was used to control for the random factor ADJECTIVE LEMMA as well as possible interactions between variables. In a second step, the effects of the constraints were compared between varieties. The resulting model predicted the choice of comparative variant correctly for 96.8 % of observations.

Results indicate that, although the adjective's number of syllables is a significant factor, its number of morphemes and frequency of use are additionally highly relevant. This result contrasts with previous studies, which determined phonological constraints to be the most important predictors of the comparative alternation (Cheung & Zhang, 2016; Hilpert, 2008). Furthermore, the effects of several

constraints are not compatible with Mondorf's (2009) *more*- support theory, thus calling into question her hypothesis that the analytic variant is easier to process. Finally, subtle variety differences indicate a possible delineation between majority-L1 and majority-L2 Englishes as well as varieties at different stages of nativization in the sense of Schneider (2007); however, the relevance of the language-internal constraints far outweighed that of the subtle variety-related differences.

References

Bernaisch, Tobias, Stefan Th. Gries, and Joybrato Mukherjee (2014), 'The Dative Alternation in South Asian English(es)', *English World-Wide* 35(1), 7–31. https://doi.org/10.1075/eww.35.1.02ber

Cheung, Lawrence, and Longtu Zhang (2016), 'Determinants of the Synthetic–Analytic Variation Across English Comparatives and Superlatives', *English Language and Linguistics* 20(3), 559–583. https://doi.org/10.1017/S1360674316000368

Davies, Mark, and Robert Fuchs (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.

Gries, Stefan Th. (2018), 'On Over- and Underuse in Learner Corpus Research and Multifactoriality in Corpus Linguistics More Generally', *Journal of Second Language Studies* 1(2), 276–308.

Heller, Benedikt, Tobias Bernaisch, and Stefan Th. Gries (2017), 'Empirical Perspectives on Two Potential Epicentres: The Genitive Alternation in Asian Englishes', *ICAME Journal* 41(1), 111–144. https://doi.org/10.1515/icame-2017-0005

Hilpert, Martin (2008), 'The English Comparative—Language Structure and Language Use', *English Language and Linguistics* 12(3), 395–417. https://doi.org/10.1017/S1360674308002694

Kytö, Merja, and Suzanne Romaine (1997), 'Competing Forms of Adjective Comparison in Modern English: What Could Be More Quicker and Easier and More Effective?', in Terttu Nevalainen and Leena Kahlas-Tarkka (eds), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, Helsinki: Société Néophilologique, 329–352.

Mondorf, Britta (2009), *More Support for More-Support: The Role of Processing Constraints on the Choice Between Synthetic and Analytic Comparative Forms*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Mondorf, Britta, and Javier Pérez-Guerra (2016), 'Special Issue on Support Strategies in Language Variation and Change', *English Language and Linguistics* 20(3), 383–393. https://doi.org/10.1017/S1360674316000289

Schneider, Edgar W. (2007), *Postcolonial English: Varieties Around the World*, Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt, and Jason Grafmiller (2023), *Comparative Variation Analysis: Grammatical Alternations in World Englishes*, Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108863742

Szmrecsanyi, Benedikt, and Bernd Kortmann (2009), 'The Morphosyntax of Varieties of English Worldwide: A Quantitative Perspective', *Lingua* 119(11), 1643–1663.

Syntax and variation in English content clauses: The case of *that* retention and omission

Veronika Hlaváčková and Gabriela Brůhová
(Charles University)

Work-In-Progress

Variation, as a defining characteristic of language, manifests prominently in the alternation of the *that*-complementiser in content clauses, which may be either retained or omitted in certain contexts (Quirk et al. 1985, Biber et al. 2021, Huddleston & Pullum 2002). Drawing on prior research (e.g., Ferreira & Dell 2000; Temperley 2003; Gadanidis et al. 2021), this study-in- progress explores syntactic and lexical factors

influencing *that*-variation, contributing to the growing understanding of language-internal factors in this domain.

The qualitative analysis adopts a corpus-based approach, utilising data from the British National Corpus (BNC). To minimize, as much as possible, extra-linguistic factors stemming from online language production, the data are restricted to written text types (e.g., books, periodicals, written-to-be-spoken texts).

A sample of 300 examples is examined for syntactic and lexical factors impacting *that*- variation. Key variables include the syntactic function of the subordinate clause (e.g., direct object, extraposed subject, subject complement) and its position in the matrix clause (medial or final). For instance, in *She decided she could safely leave him* (HA2, BNC), the content clause functions as a direct object and is positioned finally.

The analysis also investigates the embedded-clause subject, focusing on its coreference with the main-clause subject and its form (i.e. noun phrase or pronoun). In the aforementioned example, the subjects are coreferential, and the form of the embedded-clause subject is pronominal.

Additionally, the governing head of the content clause is classified lexically (e.g., *decide* being a suasive verb according to Quirk et al. 1985), and attention is also given to the competition among complementisers that could potentially substitute *that*. Such variation can be exemplified by the sentence *Check the patient understands what is to happen before preparation commences* (EV5, BNC), where the complementiser introducing the content clause can be zero, *that*, or *whether*.

In the regression analysis, each factor will be analysed both individually and collectively to determine its impact on the structure of subordinate clauses and to account for potential interactions between factors.

Preliminary findings suggest that the syntactic function of the content clause has a considerable effect on the choice between *that* omission and retention, as subordinate clauses functioning as direct objects most readily allow *that* omission. Similarly, the form of the embedded-clause subject is also significant, as pronominal embedded subjects favour omission as well. Other factors, while less significant independently, may exert influence in combination with others. The results are expected to deepen our understanding of the syntactic and lexical dynamics driving *that*-complementiser variation.

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (2021), *Grammar of Spoken and Written English*, Amsterdam: John Benjamins.

Ferreira, Victor S., and Gary S. Dell (2000), Effect of ambiguity and lexical availability on syntactic and lexical production, *Cognitive Psychology* 40(4), 296–340.

Gadanidis, Timothy, Angelika Kiss, Lex Konnelly, Katharina Pabst, Lisa Schlegl, Pocholo Umbal, and Sali A. Tagliamonte (2021), Integrating qualitative and quantitative analyses of stance: A case study of English that/zero variation, *Language in Society* 40, 1–24.

Huddleston, Rodney D., and Geoffrey K. Pullum (2002), *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985), *A Comprehensive Grammar of the English Language*, London: Longman.

Temperley, David (2003), Ambiguity avoidance in English relative clauses, *Language* 79(3), 464–484.

Sources

The British National Corpus, version 2 (BNC World), Distributed by Oxford University Computing Services on behalf of the BNC Consortium, Ústav Českého národního korpusu FF UK, Praha 2001. Available at: http://www.korpus.cz

Seriously, smugly, softly - Stage directions as explicit characterisation cues in screenplays

Christian Hoffmann
(University of Augsburg)

Stage (or screen) directions (SDs) in screenplays have long been recognised as an essential tool in the film trade and are designed to "visually foreshadow future events and reflect themes as well as characters and levels of action" (Munkelt 2004, 253). While corpus-based studies on characterisation in film and television dialogues have been gaining momentum throughout the last decade (Csomay & Young 2021; Werner 2021; Bednarek 2023), linguistic work that shows how SDs contribute to characterisation in theatre scripts or film screenplays is rare. Given this research lacuna, Locher et al (2023, 17) have recently called for a more systematic exploration of SDs to systematically map their formal and functional patterns across larger sets of scripts or screenplays. This corpus-based study examines the lexical patterns that shape SDs in a corpus of 100 different US Hollywood screenplays, balanced and annotated for production decades, film genres, character types (main or supporting characters) and character gender (male, female). The goal is to examine the relative frequency and dispersion of adjectives, adverbs and prepositional phrases in two of the most common SD text segments, namely *parentheticals* and *actions texts*. While parentheticals specify the spatial, emotive or prosodic quality of ensuing character speech, action texts contain more elaborate descriptions of characters' mental or physical state, appearance, position, or actions at the outset of any new film scene. The data used for the study was sampled from a range of different online screenplay databases, normalized for spelling variations, cleaned of metainformation, and tagged for parts of speech using CLAWS (Garside & Smith 1997). Python scripts allowed me to clean and segment the resulting SD data and assign the systematic mark up. For parentheticals, relevant lexical units were accessed from the different sub corpora by queries in *Antconc* (Lawrence 2024). In contrast, action texts were first aggregated for different film genres. Each genre dataset then constituted its own target corpus, with all remaining genre datasets serving as the corresponding reference corpus. The setup was used to conduct a key POS analysis, using *WMatrix* (Rayson & Smith 2006). The basic assumption tested is that if the lexical patterns show similar frequencies and have similar distributions across the datasets, it yields a set of highly conventionalized SD expressions that screenwriters routinely rely on for characterization purposes. If, however, lexical patterns prove to be dispersed unevenly across the datasets, this might indicate that:

(a) screenwriters rely on diverging gender representations for fictional characters,
(b) different film genres privilege different types of character descriptions, and
(c) the use of lexical patterns in SDs has evolved from the 1970s to the present day.

The findings of the study do not only present a valuable addition to existing corpus stylistic research on fictional characterization but arguably have a practical merit for screenwriters whose lexical choices can now be put on a proper empirical footing.

References

Anthony, Lawrence (2024), AntConc (Version 4.3.1), Tokyo: Waseda University. https://laurenceanthony.net/software/antconc/

Bednarek, Monika (2023). *Language and Characterisation in Television Series*. *A corpus-informed approach to the construction of social identity in the media,* Amsterdam: Benjamins.

Csomay, Eniko and Ryan Young (2021), Language use in pop culture over three decades: A diachronic keyword analysis of Star Trek dialogues, *International Journal of Corpus Linguistics* 26(1), 71–94.

Garside, Roger and Nicolas Smith (1997), A hybrid grammatical tagger: CLAWS4, in Roger Garside, Geoffrey Leech & Tony McEnery (eds), *Corpus annotation: Linguistic information from computer text corpora*, London: Longman, 102–121.

Locher, Miriam A., Andreas H. Jucker, Daniela Landert, and Thomas C. Messerli (2023), *Fiction and Pragmatics,* Cambridge: Cambridge University Press.

Munkelt, Marga (1987), Stage directions as part of the text, *Shakespeare Studies 19,* 253–272.

Rayson, Paul and Nicholas Smith (2006), The key domain method for the study of language varieties, *The Third Inter Varietal Applied Corpus Studies (VACS) Group International Conference on "Language At The Interface*", University of Nottingham, UK, 23–24 June 2006.

Rowen, Bess (2018), Undigested Reading: Rethinking Stage Directions through Affect, *Theatre Journal 70* (3), 307–326.

Werner, Valentin (2021), A diachronic perspective on telecinematic language, *International Journal of Corpus Linguistics* 26(1), 38–70.

Writing trajectories of grammatical complexity in L1 children's writing

Christian Holmberg Sjöling and Taehyeong Kim
(Luleå University of Technology, Northern Arizona University)

Children have typically grasped basic concepts of English grammar when they begin school, and as they go through formal education, their ability to express themselves with complexity and contextual sensitivity in writing develops (Applebee, 2000; Hoff, 2009). The written language pupils produce changes as they mature since they, for example, learn to express a wider range of ideas through different types of texts intended for different audiences (Rose & Martin, 2012). In this paper, we aim to add to our cumulative knowledge of how this process develops by automatically analysing the development of a wide range of grammatical complexity features in L1 children's writing across different years of education in relation to Biber et al's (2011) developmental sequence (cf. Durrant et al., 2020; Durrant & Brenchley, 2022). The analysed material consists of 884 texts from the *Growth in Grammar* corpus (Durrant, 2019) written by L1 pupils across Britain between 2015 and 2017 as a part of their regular schoolwork for years 2, 4, 9, and 11. We seek to answer the following questions:

1. How does the use of grammatical complexity features develop across different years of education?
2. To what extent does the grammatical complexity of students' writing in their final year of education resemble that of advanced writers?

To answer these questions, we rely on the Descriptive Grammar Approach (Biber et al., 2024) by computing and analysing grammatical complexity features that are linguistically interpretable (e.g., frequency of finite verb complement clause *that* and nouns as nominal pre- modifiers averaged per 1,000 words). The automated analysis is carried out with the newly released *Lexicogrammatical Tagger* (LxGrTgr) (Kyle, 2025). In order to focus on linguistic interpretation, we employ a minimally sufficient statistical approach (see Larsson et al., 2022) for analysing trajectories which goes beyond mean rates of occurrence and inferential statistics (Staples et al., 2023). It includes, amongst other things, conducting Spearman's correlations followed by visual inspection of the plotted data. The results show clear developmental trends for several grammatical complexity features as pupils gradually produce more non-finite verb complement clauses (*to*) (e.g., the marriage is shown **to be entirely loveless**), attributive adjectives (e.g., **cheerful** vibes), *of* phrases as postmodifiers (e.g., The taste **of** freedom slips away), and finite noun + *that* complement clauses (e.g., it is also linked to **the fact that** Lady Macbeth wants her femininity). There is also a decreasing trend for premodifying nouns (e.g., A few days later naughty **diamond** fairies). Taken together, the findings show that children's written production of grammatical complexity features increase as years of education increase. Texts written by students during their final year of education also resemble those produced by more advanced writers as they include features characteristic of academic writing (e.g., appositives and non-finite complement clauses (*to*). Lastly, we discuss our results in terms of how they can help curriculum developers, practitioners, and test developers improve children's writing development.

References

Applebee, Arthur N. (2000), Alternative models of writing development, in R. Indrisano and J. Squire, R. (eds), *Perspectives on Writing: Research, theory and practice,* International Reading Association, 90–110.

Biber, Douglas, Bethany Gray, and Kornwipa Poonpon (2011), Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly* 45(1), 5–35.

Biber, Douglas, Tove Larsson, Gregory R. Hancock, Randi Reppen, Shelley Staples, and Bethany Gray (2024), Comparing theory-based models of grammatical complexity in student writing, *International Journal of Learner Corpus Research* 11(1), 145–177.

Durrant, Philip (2019), *Growth in grammar corpus 2015-2019,* [Data Collection], Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-853809

Durrant, Philip, Mark Brenchley, and Rebecca Clarkson (2020), Syntactic development across genres in children's writing: the case of adverbial clauses, *Journal of Writing Research* 12(2), 419–52.

Durrant, Philip and Mark Brenchley (2023), Development of noun phrase complexity across genres in children's writing, *Applied Linguistics* 44(2), 239–264.

Hoff, Erika (2009), *Language Development,* Wadworth, Cengage Learning.

Kyle, Kristopher (2025), Lexicogrammatical tagger (LxGrTgr), GitHub. GitHub -kristopherkyle/LxGrTgr: Lexicogrammatical Tagger

Larsson, Tove, Jesse Egbert, and Douglas Biber (2022), On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective, *Corpora* 17(1), 137–157.

Rose, David and James R. Martin (2012), Learning to Write/ Reading to Learn: Genre, Knowledge and Pedagogy in the Sydney School, *Language and Education* 28(1), 100–101, Equinox.

Staples, Shelley, Bethany Gray, Douglas Biber, and Jesse Egbert (2023), Writing trajectories of grammatical complexity at the university: comparing L1 and L2 English writers in BAWE, *Applied Linguistics* 44(1), 46–71.

To write etceteraly: The use of 'et cetera' abbreviations in Early English medical corpora

Alpo Honkapohja
(Tallinn University)

Abbreviations for the Latin phrase *et cetera* are among the most long-standing abbreviation conventions in the English language. They hold a particular association with scientific writing. For instance, *Blackwood's Edinburgh Magazine* (1822) humorously critiqued the phrase's overuse in scientific discourse, coining the adverbial form 'etceteraly' to satirize its ubiquity: 'To write critically, scientifically.. etceteraly' (OED). The present paper investigates the functions, contexts, and evolution of this familiar discourse-organising phrase was used in Middle English and Early Modern English medical texts.

Adopting a historical pragmatic approach, the study employs a classic form-to-function mapping (cf. Jucker 1995) to analyse the functions of *et cetera* across the medical register. I will work with two historical corpora as the dataset: *Middle English Medical Texts* (MEMT, 2005) and *Early Modern English Medical Texts* (EMEMT, 2010). Using AntConc, I will identify all spelling variants of 'et cetera', including both contracted (*&c. & c´*) and non-contracted forms (*et caetera*, *et cetera*, *etcetera*), which are then analyzed in contexts using KWIC to assign them to various subcategories. These include occurrences at the end of a recipe, at the end of a text, and at the end of a prayer. The latter can be identified by looking at collocates of 'Amen' since it was conventional to omit well known prayers and invocations to the Lord. The results will also be examined across the text categories in the corpora, such as surgeries, *materia medica*, and specialized treatises in MEMT; and scientific journals, general treatises or textbooks, texts on specific diseases, methods, substances, midwifery and plague, recipe collections and regiments in EMEMT. In addition, the diachronic aspect of the EMEMT corpus, with its more precise publication dates, facilitates a comparative

analysis over time. The results will be analyzed both quantitatively, to establish the frequency of *et cetera* in different contexts, and qualitatively, to explore its shifting pragmatic functions.

Research question:

1. What were the primary pragmatic functions of *et cetera* in Middle English versus Early Modern English medical texts?
2. How does the frequency and distribution of *et cetera* vary across medical genres and text types within the MEMT and EMEMT corpora?
3. To what extent does the shift from manuscript to print culture influence the form and function of *et cetera*?
4. Can the development of *et cetera* in medical texts be linked to wider trends in scientific discourse and standardization during the period?

The results are expected to reveal a broadening of the pragmatic functions of 'et cetera', spreading into new contexts, including the development toward less formulaic uses. The texts included in MEMT all originate in manuscripts, in which saving paper and parchment was a major concern. However, 'et cetera' has proven to be a flexible phrase which has been in use in all periods in the History of English.

References

Jucker, Andreas H. (ed) (1995), *Historical Pragmatics: Pragmatic Developments in the History of English*, Amsterdam and Philadelphia: John Benjamins.

Oxford English Dictionary (OED), Oxford: Oxford University Press.

Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr, and Jukka Tyrkkö (compilers) (2010), *Corpus of Early English Medical Writing*. Electronic resource published together with I. Taavitsainen and P. Pahta (eds), *Early Modern English Medical Texts: Corpus Description and Studies*, Amsterdam and Philadelphia: John Benjamins. Available at: http://doi.org/10.1075/z.160

The effects of Extramural English reading on phraseology in L2 writing: A key phrase frames approach

Taehyeong Kim, Tove Larsson, Henrik Kaatari, Ying Wang and Pia Sundqvist
(Northern Arizona University, University of Gävle, Karlstad University, University of Oslo)

Second language (L2) learners worldwide are increasingly exposed to English outside the classroom through self-initiated, Extramural English (EE) activities (Sundqvist, 2009). EE has attracted particular attention from researchers and educators in countries like Sweden, where English is a foreign language and yet well-integrated into daily life. The few studies that have explored the effect of EE on L2 writing have shown that frequent EE engagement leads to the use of more diverse (Kaatari et al., 2023) and less frequent (Olsson, 2016) vocabulary. EE reading stands out as a particularly important activity for students to engage in. Previous studies have demonstrated that EE reading fosters knowledge of recurrent word combinations such as phrasal verbs (Garnier & Schmitt, 2016), collocations (González-Fernández & Schmitt, 2015), and adjective-noun combinations (Wang et al., 2024). Building on these findings, the present study adopts a frequency-based approach to examine whether the effect of language input from EE reading extends beyond word pairs, to longer multi-word units, specifically to discontinuous four-word sequences with a variable slot (e.g., *the most * aspect*), known as *phrase frames* (*p*-frames). The variable slot is filled by words referred to as *fillers*. Previous studies have shown that high variability among fillers and structural characteristics of *p*-frames involving phrasal complexity are important indicators of advanced L2 production (e.g., Garner, 2016; Larsson et al., 2022; Tan & Römer, 2022).

The present study used the Swedish Learner English Corpus (SLEC; Kaatari et al., 2024), which comprises L2 English junior and senior high school writing, to identify groups of learners who read in English every week versus those who do not. We then applied a novel method taking text dispersion into

consideration to identify *p*-frames that are *key* to the reading and non-reading groups in reference to each other, respectively. Based on these key *p*-frames, our study asks:

- To what extent do the reading and non-reading groups differ in terms of variability of fillers in key *p*-frames?
- To what extent do the reading and non-reading groups differ in terms of structural characteristics of key *p*-frames?

The results show that key *p*-frames of the reading group are characterized by high variability, a frequent use of post-nominal modifiers (e.g., *the * of a*), and lower frequencies of personal pronouns, that is, features associated with informational writing (e.g., Biber, 2006). In contrast, the non-reading group used key *p*-frames that are characterized by lower variability; they also included more embedded clauses and personal pronouns, that is, features common in more involved production (e.g., Biber, 2006). Pedagogical implications for the development of phraseology in L2 writing are discussed, highlighting the role of language exposure through self-initiated reading.

References

Biber, Douglas (2006), Stance in spoken and written university registers, *Journal of English for academic purposes* 5(2), 97–116. https://doi.org/10.1016/j.jeap.2006.05.001

Garner, Jamie (2016), A phrase-frame approach to investigating phraseology in learner writing across proficiency levels, *International Journal of Learner Corpus Research* 2(1), 31–67. https://doi.org/10.1075/ijlcr.2.1.02gar

Garnier, Mélodie, and Norbert Schmitt (2016), Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System* 59, 29–44. https://doi.org/10.1016/j.system.2016.04.004

González Fernández, Beatriz and Norbert Schmitt (2015), How much collocation knowledge do L2 learners have? The effects of frequency and amount of exposure, *ITL-international journal of applied linguistics* 166(1), 94–126.

Kaatari, Henrik, Tove Larsson, Ying Wang, Seda Acikara-Eickhoff, and Pia Sundqvist (2023), Exploring the effects of target-language extramural activities on students' written production, *Journal of Second Language Writing* 62, 101062.

Kaatari, Henrik, Ying Wang, and Tove Larsson (2024), Introducing the Swedish Learner English Corpus: A corpus that enables investigations of the impact of extramural activities on L2 writing, *Corpora* 19(1), 17–30. https://doi.org/10.3366/cor.2024.0296

Larsson, Tove, Randi Reppen, and Tülay Dixon (2022), A phraseological study of highlighting strategies in novice and expert writing, *Journal of English for Academic Purposes* 60, 101179. https://doi.org/10.1016/j.jeap.2022.101179

Olsson, Eva (2016), *On the impact of extramural English and CLIL on productive vocabulary* [Doctoral dissertation, University of Gothenburg], Gupea. https://gupea.ub.gu.se/handle/2077/41359

Sundqvist, Pia (2009), *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary* (Karlstad University Studies, 2009:55) [Doctoral dissertation, Karlstad University]. DiVA. https://www.diva-portal.org/smash/get/diva2:275141/FULLTEXT03.pdf

Tan, Yi, and Ute Römer (2022), Using phrase-frames to trace the language development of L1 Chinese learners of English. *System* 108, 102844. https://doi.org/10.1016/j.system.2022.102844

Investigating historical fiction in the TV Corpus

Catherine Laliberté
(Ludwig Maximilian University of Munich)

Work-In-progress

From a linguistic perspective, historical fiction is a remarkable genre. Representing the past in literary or telecinematic fiction typically requires making stylistic choices to convey a sense of "old-timey"-ness. Insofar as it builds imagined worlds, historical fiction is not unlike science fiction and fantasy in its use of estranging linguistic devices (cf. Adams 2017). Writers of historical fiction such as Hilary Mantel (2013: 136), David Mitchell (cf. Stocker 2012), and Julian Fellowes (Friedlander 2024) speak of their craft as a balancing act between accuracy (being true-to-the-record), authenticity (being believable, cf. Saxton 2020), and being mindful of using language their audience can understand and appreciate. Among other things, the fact that writers share this common perspective indicates that historical fiction as a whole is linguistically special and can be systematically investigated using corpus methods.

The present contribution uses corpus methods to investigate what makes the language of historical fiction distinct from other types of fiction. The study focuses on televisual historical fiction, an entertainment genre that is currently extraordinarily popular (e.g. Ingram 2023; Peters & Sperling 2023). Studies on individual TV series such as *Bridgerton* (Laliberté, Keller & Wengler 2024) and *Downton Abbey* (Bruti & Vignozzi 2016) have hinted at features such as the high frequency of *must*, the particular prominence of forms of address (e.g. *Lady*, *Ladyship*), and the relative lack of contractions as being representative of the genre. This study aims to analyze this type of fiction on a larger scale, using the TV Corpus (Davies 2019) and the historical fiction dialogue it contains, which amounts to some 3,000 texts and 15 million words. In addition to testing the importance of some of the features named above, the present study will analyze parts-of-speech (POS) and semantic categories with the help of Wmatrix (Rayson 2009), as such analyses have been previously successfully employed to expose genre-specific stylistic features (see Castro 2023). Incidentally, the findings may have methodological implications: capturing the specificities of telecinematic historical fiction has the potential to highlight representativity issues that may come with the inclusion of historical fiction material in generalist corpora, as brought up by Egan (2019).

This study is part of a larger project investigating the role of television in the public's understanding of the past and, in particular, the history of English. Television is the main vehicle for the packaging of history for the general public and plays a significant role in shaping people's understanding of their past (cf. de Groot 2008). Ultimately, describing stylized "historical" language illuminates how beliefs about English and its history are formed, perpetuated, and interrogated.

References

Adams, Michael (2017), The pragmatics of fantasy and science fiction, in M. A. Locher, and A. H. Jucker (eds), *The Pragmatics of Fiction*, Berlin: De Gruyter, 329–363.

Bruti, Silvia, and Gianmarco Vignozzi (2016), Routines as social pleasantries in period dramas: A corpus linguistics analysis, in S. Bruti, and R. Ferrari (eds), *A Language of One's Own*, Bologna: I libri di Emil, 207–240.

Castro, Adrián (2023), Telecinematic stylistics: Language and style in fantasy TV series, *Language and Literature* 33(1), 3–24.

Davies, Mark (2019), *The TV Corpus*. Available at: https://www.english-corpora.org/tv/ (accessed 22 November 2024).

De Groot, Jerome (2008), *Consuming History*, London: Routledge.

Egan, Thomas (2019), Non-representativeness in corpora: Perils, pitfalls and challenges, *CogniTextes* 19.

Friedlander, Whitney (2024), How *The Gilded Age*, *Feud* and more period dramas walk a fine line between being accurate and offensive with language, *Variety*, 31 May. Available at:

https://variety.com/2024/tv/awards/gilded-age-feud-period-dramas-language-1236018015/ (accessed 26 November 2024).

Ingram, Hunter (2023), From *Bridgerton* to *Shōgun*, television looks to history for 2024's biggest shows, *TV Guide*, 27 December. Available at: https://www.tvguide.com/news/2024-tv-shows-historical-dramas/ (accessed 26 November 2024).

Laliberté, Catherine, Melanie Keller, and Diana Wengler (2024), Linguistic strategies of estrangement in historical fiction: *Bridgerton* and *Downton Abbey*, *Anglistik* 35(3).

Mantel, Hilary (2013), Untitled, in C. Brayfield, and D. Sprott (eds), *Writing Historical Fiction: A Writers' and Artists' Companion*, London: Bloomsbury, 135–136.

Peters, Jeremy W., and Nicole Sperling (2023), 'Based on a true story' (except the parts that aren't), *The New York Times*, 14 January. Available at: https://www.nytimes.com/2023/01/14/business/media/tv-historical-dramas-fictional.html (accessed 26 November 2024).

Rayson, Paul (2009), *Wmatrix: A Web-Based Corpus Processing Environment*, Computing Department, Lancaster University. Available at: http://ucrel.lancs.ac.uk/wmatrix/ (accessed 27 November 2024).

Saxton, Laura (2020), A true story: Defining accuracy and authenticity in historical fiction, *Rethinking History* 24(2), 127–144.

Stocker, Bryony D. (2012), 'Bygonese' – Is this really the authentic language of historical fiction? *New Writing* 9(3), 308–318.

Spontaneity and fiction in spoken language: Analysing dialogues in the Improv Corpus

Daniela Landert and Lea Kyveli Chrysanthopoulou
(Heidelberg University)

Spoken language is far from homogeneous across different contexts. However, in comparison with written language, the knowledge about how contextual factors affect the use of spoken language is still rather limited. Two aspects that have often been conflated in previous research are, on the one hand, whether spoken language is spontaneously produced or scripted and, on the other, whether spoken language is used in a work of fiction (TV series, movies) or in a non-fiction setting. Previous research has been able to show that there are differences between spontaneous non-fiction conversation and scripted dialogues in fiction (e.g. Bednarek 2018; Bublitz 2017; Jucker 2021; Quaglio 2009), but whether these differences are due to the spontaneous nature of language production or due to the functional demands of language in fiction remains an open question in many cases.

In our study, we address this issue with the help of a new corpus that consists of spoken language in spontaneously produced fiction. The Improv Corpus is based on 50 video recordings of improvised theatrical fiction, which are manually transcribed. The corpus composition is still in progress and for this study we base our result on a subset of the 28 recordings that are already transcribed, which cover almost 22 hours of recordings and amount to 206,000 words. We compare this data set to scripted dialogues from TV series, which were also manually transcribed, as well as to spoken language in non-fiction settings, based on the Santa Barbara Corpus. For the Santa Barbara Corpus, we distinguish between spontaneous conversation, task-based interaction, and non-dialogic spoken language, such as lectures and sermons. This results in two sets of spoken language from fiction (scripted and spontaneous) and three sets of non-fictional spoken language with different degrees of spontaneity. Based on these five data sets, we analyse linguistic characteristics that have been associated with spontaneous language production, such as overlaps, hesitations, false starts, and discourse markers, as well as more general formal and structural characteristics, such as the distribution of parts-of-speech and the use modal verbs and pronouns. We combine automatic processing of the corpus data with the Natural Language Toolkit (Bird et al. 2009) and an adaptation of Burrow's Delta (Burrows 2002), with a more detailed qualitative analysis of selected

features. We interpret these results from a pragmatic perspective, focusing on how the contextual constraints of language production in each setting shape the use of linguistic features.

Our results show that while there are features that correlate with either spontaneity or fiction, the situation is more complex in many cases. Especially, the link between spontaneous language production and linguistic features associated with spontaneity is less clear than expected. Thus, we suggest that it is time for a reassessment of the notion of spontaneity in spoken language and that additional dimensions of spoken language production should be explored in more detail.

References

Bednarek, Monika (2018), *Language and Television Series: A Linguistic Approach to TV Dialogue*, Cambridge: Cambridge University Press.

Bird, Steven, Ewan Klein, and Edward Loper (2009), *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, Sebastopol, CA: O'Reilly Media.

Bublitz, Wolfram (2017), Oral features in fiction. In: Locher, Miriam A., and Andreas H. Jucker (eds.), *Pragmatics of Fiction*, Berlin/Boston: De Gruyter Mouton, 235–263.

Burrows, John F. (2002), Delta: A measure of stylistic difference and a guide to likely authorship, *Literary and Linguistic Computing* 17(3), 267–287.

Jucker, Andreas H. (2021), Features of orality in the language of fiction: A corpus-based investigation, *Language and Literature* 30(4), 341–360.

Quaglio, Paulo (2009), *Television Dialogue: The Sitcom Friends vs. Natural Conversation*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

"God save the Queen and her language!" A corpus-based attitudinal study of Indian English newspaper language

Claudia Lange and Sven Leuckert
(University of Dresden)

By design, the *International Corpus of English* (ICE) project focuses on national standard(ising) varieties of English (Greenbaum 1990), and the availability of by now more than a dozen ICE components has given a tremendous boost to empirical research on World Englishes. So far, the only ICE-component to factor in regional differentiation is ICE-Ireland, comprising North and South Ireland and thus straddling the national boundary between the UK and the Republic of Ireland (Kallen & Kirk 2008). Other corpus projects have been created to address regional and/or ethnic variation within national varieties, such as a range of corpora sampling Black South African English (van Rooy 2013), or the *Corpus of Regional Indian Newspaper Englishes* (CORINNE, cf. Yurchenko et al. 2021) for India. CORINNE has already been used to investigate the regional distribution of a range of morphosyntactic and lexical features of Indian English across five Indian regions which display different degrees of the entrenchment of English in their societies (see Leuckert et al. 2023). This is in line with Sharma (2023), who stresses that Indian English is not a monolithic entity across the subcontinent.

This paper extends the regionally differentiated perspective on Indian English(es) by using CORINNE, a 27-million-word corpus of newspaper language, as the basis for an attitudinal study. We ask whether the specific social, historical, and political configurations which impact the Indian regional states' language policies and thus the role of English within their respective communicative spaces (cf. Leuckert at al. 2023: 3-15) have a linguistic correlate. More specifically, the question is whether the semantic prosody of a set of *language*-related keywords drawn from the regional CORINNE-subcorpora aligns with that region's public stance and/or policy towards English as an "aspirational language" (Graddol 2010: 64), Hindi as official and potential national language, and the status of local and minority languages. In order to address this question, we extracted the Keyword-In-Context (KWIC) concordance lines for *language* (with modifiers *national*, *regional*, *state*, *official*), *ras(h)tra bhasha* ('national language'), and *vernacular*,

and investigated the resulting sub-corpora using (a) the diversity/unity (d/u) ratio described in Mukherjee and Bernaisch (2015: 421), which was previously applied to national as opposed to regional varieties, (b) sentiment analysis (see, for instance, Lei & Liu 2021), and (c) collexeme analysis (see Tang 2016).

Preliminary results show that the regional newspapers differ in terms of their expressed attitudes towards language(s) (particularly English), with the d/u ratio showing significant differences between Indian states. While attitudes are never exclusively negative, sentiment analysis, again, reveals regional differences. The collexeme analysis, finally, shows typical constructions in which the keywords appear, which frequently involve other languages (such as English, Hindi, or local languages), but also highly emotionally charged words. Overall, the results indicate that a regional perspective on language use in India is not only needed for the investigation of structural differences at a higher level of granularity, but may also prove highly insightful for the study of language attitudes.

References

Bernaisch, Tobias, Benedikt Heller, and Joybrato Mukherjee (2021), *Manual for the 2020-Update of the South Asian Varieties of English (SAVE2020) Corpus*, Version 1.1, Giessen: Justus Liebig University, Department of English.

Graddol, David (2010), *English Next India: The Future of English in India*, New Delhi: British Council.

Greenbaum, Sidney (1990), 'Standard English and the International Corpus of English', *World Englishes* 9(1), 79–83.

Kallen, Jeffrey L., and John M. Kirk (2008), *ICE-Ireland: A User's Guide. Documentation to Accompany the Ireland Component of the International Corpus of English (ICE-Ireland)*, Belfast: Cló Ollscoil na Banríona.

Lei, Lei, and Dilin Liu (2021), *Conducting Sentiment Analysis*, Cambridge: Cambridge University Press.

Leuckert, Sven, Asya Yurchenko, Claudia Lange, and Tobias Bernaisch (2023), *Indian Englishes in the Twenty-First Century: Unity and Diversity in Lexicon and Morphosyntax*, Cambridge: Cambridge University Press.

Mukherjee, Joybrato, and Tobias Bernaisch (2015), 'Cultural Keywords in Context – A Pilot Study of Linguistic Acculturation in South Asian Englishes', in Peter Collins (ed.), *Grammatical Change in English World-Wide*, Amsterdam: John Benjamins, 411–435.

Sharma, Devyani (2023), *From Deficit to Dialect: The Evolution of English in India and Singapore*, Oxford: Oxford University Press.

Tang, Xuri (2016), 'Lexeme-Based Collexeme Analysis with DepCluster', *Corpus Linguistics and Linguistic Theory*, 13(1), 165–202.

van Rooy, Bertus (2013), 'Corpus Linguistic Work on Black South African English: An Overview of the Corpus Revolution and New Directions in Black English Syntax', *English Today*, 29(1), 10–15.

Yurchenko, Asya, Sven Leuckert, and Claudia Lange (2021), 'Comparing Written Indian Englishes with the New Corpus of Regional Newspaper Englishes (CORINNE)', *ICAME Journal*, 45(1), 179–205.

Who interrupts? Investigating contestive and supportive interruptions in FOMC meetings

Christian Langerfeld and Gisle Andersen
(Norwegian School of Economics)

Meeting interaction is among the most essential forms of discourse in organisational life, and this paper builds upon earlier research on a central topic in meeting interaction, namely interruptions and overlapping speech. We focus on the high-stakes context of meetings held by the Federal Open Market Committee in the USA. For the study, we use the freely available FOMC corpus introduced by Langerfeld and Andersen (2023, 2024). The corpus comprises approximately 11,000,000 words from transcripts of the meetings of the U.S. Federal Open Market Committee, spanning the years 1987 to 2018.

Taking a discourse-analytical and socio-pragmatic approach, we aim in our presentation to explore a set of issues related to the dynamics of interruptions by asking questions about the interruptions themselves as well as which participants engage as interrupter and interrupted.

The dataset has been enriched with metadata on the participants according to several variables. This allows for specific corpus-based searches and corpus-driven statistics on speaker criteria, and thus enables us to explore the question posed in the title with regard to factors such as speaker gender, age and position in the organisational hierarchy (Chair/ Delegate, i.e. non-chair participant). The study focuses on how these variables interact and contribute to the overall patterns of variability. As is well known, interruptions can have supportive and contestive functions, and we therefore also aim to look for discernible differences between groups of speakers as to whether interruptions seem to be primarily contestive or supportive (such as barging in to present own view, vs. joint construction of ideas, respectively).

Our preliminary quantitative analysis of the FOMC corpus suggests that female meeting participants are not interrupted more frequently than their male colleagues. This stands in contrast to much previous research on interruptions in meeting discourse, e.g., the findings by Van Eecke and Fernández (2016). Our results show that, on average, male participants are interrupted more often than female participants within our sample. Furthermore, our analysis reveals a noteworthy pattern at the individual level: participants who frequently interrupt others also tend to be interrupted more often themselves. The relationship between being an interrupter and being an interrupted is statistically significant. This finding appears to be independent of gender and may be indicative of individual meeting behaviours and the dynamics of the meeting as a whole.

Aries (1996) suggests considering other factors beyond gender, such as power dynamics when investigating interactional behaviour. Power relations can manifest themselves in various ways. The corpus enables us to examine whether the interactional behaviour of meeting chairs – who occupy the highest position in the meeting hierarchy – differs from that of delegates with respect to interruptions. Our analysis aims to discover more subtle patters of variability, including whether age plays any role at all, and whether females are more supportive and males more contestive in their interruptions, as suggested in previous literature on organisational discourse.

References

Andersen, Gisle and Christian Langerfeld (2024), Humour and laughter as indicators of meeting leadership style in FOMC meetings, *Discourse Studies,* 0(0). Available at: https://doi.org/10.1177/14614456241276762

Aries, Elizabeth (1996), *Men and Women in Interaction: Reconsidering the Differences*, New York: Oxford University Press. Available at: https://doi.org/10.1093/oso/9780195094695.001.0001

Langerfeld, Christian and Gisle Andersen (2023), The Dynamics of Turn-taking in Meetings of the Federal Open Market Committee, *Fachsprache. Journal of Professional and Scientific Communication* 45(3-4), 187–210.

Van Eecke, Paul and Raquel Fernández (2016), On the influence of gender on interruptions in multiparty dialogue, in *Proceedings of Interspeech 2016*, 2070–2074. Available at: https://doi.org/10.21437/Interspeech.2016-951

Orthographic evidence for Older Scots long front vowel changes

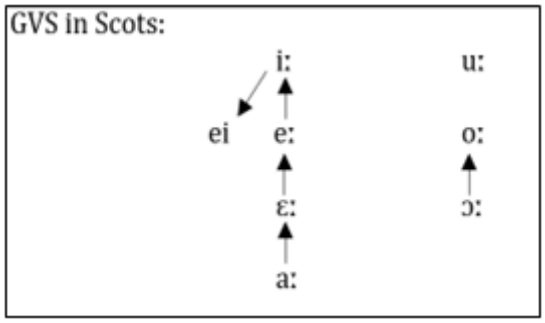
Pia Lehecka
(University of Edinburgh)

This paper explores changes in spelling practices for representations of Older Scots long front vowels /iː/, /eː/, /ɛː/, /aː/, and how this data can be used to reconstruct sound changes affecting these vowels. Like Middle English, Older Scots underwent a series of long vowel changes, including /iː/ diphthongising to /ei/ then /ai/, visible when comparing early Middle English (eME) /siːn/ to Modern Scots (MdSc) /səin/

since. Early ME /eː/ and /ɛː/ raised and merged at /iː/, e.g. eME /seː/ *see* and /sɛː/ *sea* both result in MdSc /siː/. Early ME /aː/ raised to /eː/, as in eME /maːr/ to MdSc /meːr/ *more*. These Scots long-vowel changes and the Great Vowel Shift (GVS) have been approximately dated by scholars such as Johnston (1997) and Aitken & Macafee (2002). However, a systematic corpus-based analysis of spelling evidence to date the changes has not yet been done. Exploring the development of long vowels in Scots not only has merit in its own right, but allows us to assess whether the GVS was a pull chain or a push chain (see Figure 1; cf. Jespersen (1909), Stockwell & Minkova (1988), Luick (1896), and Lass (2000)).

Data

The orthographic data for this research is drawn from three corpora. *The Helsinki Corpus of Older Scots* (HCOS, 1450-1700) is untagged and searchable by strings, and consists of 71 varied genre texts of 834,200 tokens. *The Linguistic Atlas of Older Scots* (LAOS, 1375-1500) is lexically-grammatically tagged from 1,250 local administrative texts of c.400,000 tokens. Both these corpora require extensive data extraction to quantitatively analyse spellings of Older Scots vowels. The key resource of my research is LAOS's sub-corpus *From Inglis to Scots* (FITS), which is additionally grapho-phonologically tagged and consists of c.110,000 tokens. FITS phonologically reconstructs Older Scots by triangulating between spelling, sound changes, origin of the item and present-day pronunciation. This grapho-phonological mapping creates a corpus of extensive spelling convention profiles of all sounds, visually presented by *Medusa* (Figure 2). The data resulting from FITS allows immediate quantitative analysis of the development of specific phonemes across various phonotactic and grammatical contexts, including localising sound changes via spelling changes in time and space.



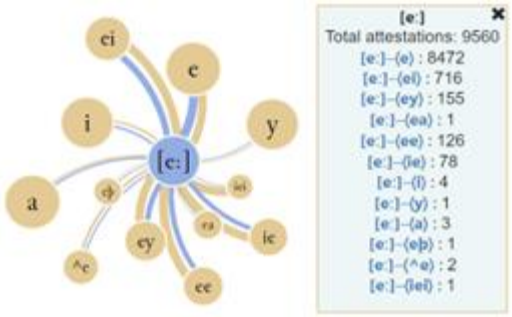


Figure 1: Scots Great Vowel Shift Figure 2: *Medusa* visualisation of [eː] graphemes in FITS

Method

This paper investigates how spelling practices representing etymological /iː/, /eː/, /ɛː/ and /aː/ change between 1375 and 1700 and how this orthographic data can refute, or support and date these vowel shifts. This is done by tracing the (relative) frequency of spelling variants of these vowels across the entire Older Scots period and by studying isolate and unconventional spellings in detail. Orthographic research of this kind follows the *littera* approach, which is based on the theory that sound shifts are reflected in spelling changes in a pre-standard variety (Laing 1999, Laing and Lass 2003). Spelling changes are the primary evidence of sound changes in this time period and this project is the first to quantitatively investigate orthography and the GVS in Older Scots in corpora.

Preliminary and expected results

Results of FITS, LAOS and HCOS data analysis show innovative <yi> digraph spellings suggesting first diphthongisation of /iː/ in the first half of the 15th ct. in spellings such as <vyis> *wise*. Widening of the diphthong is suggested by <ei> and <ai>-type spellings, which emerge and increase in the early 16th ct., e.g. <cheild> *child*, <faine> *fine*. The raising of /eː/ to /iː/ is reflected in a new innovation of <i>-type spellings, e.g. <stieire> *steer* in the second half of the 15th ct. For both high vowels, these innovative spellings increase until 1600, after which they rapidly decrease and disappear by 1700. Innovative spellings of /ɛː/ suggesting raising are absent until 1500. <e>-type spellings indicating raising of /aː/ to /ɛː/ begin to emerge in 1430 and consistently occur throughout the remainder of the 15th ct., e.g. <sem> *same*, <heile> *whole*. Expected

results of the HCOS analysis include a continuation and increase of the <e>-type spelling innovation in /aː/ items, and an emergence of isolate <i>-type spellings representing etymological /ɛː/.

References

Aitken, Adam Jack, and Caroline Macafee (2002), *The Older Scots Vowels: A History of the Stressed Vowels of Older Scots from the Beginnings to the Eighteenth Century*, Edinburgh: Scottish Text Society.

Jespersen, Otto (1909), *A Modern English Grammar on Historical Principles. Part I: Sounds and Spellings*, Heidelberg: Carl Winter's Universitätsbuchhandlung.

Johnston, Paul (1997), 'Older Scots Phonology and Its Regional Variation', in Charles Jones (ed.), *The Edinburgh History of the Scots Language*, Edinburgh: Edinburgh University Press, 47–111.

Laing, Margaret (1999), 'Confusion wrs confounded: Litteral Substitution Sets in Early Middle English Writing Systems', *Neuphilologische Mitteilungen*, 100(3), 251–270.

Laing, Margaret, and Roger Lass (2003), 'Tales of 1001 Nists: The Phonological Implications of Litteral Substitution Sets in Some Thirteenth-Century South-West Midland Texts', *English Language and Linguistics*, 7(2), 257–278.

Lass, Roger (2000), 'Phonology and Morphology, III', in Roger Lass (ed.), *The Cambridge History of the English Language*, vol. 3, Cambridge: Cambridge University Press, 56–186. https://doi.org/10.1017/chol9780521264761.004

Los, Bettelou, Rhona Alcorn, Vasileios Karaiskos, Warren Maguire, Joanna Kopaczyk, Benjamin Molineaux, and Daisy Smith (2014–2018), *FITS – From Inglis to Scots: Mapping Sounds to Spellings*, Edinburgh: University of Edinburgh, School of Philosophy, Psychology and Language Sciences.

Luick, Karl (1896), *Untersuchungen zur Englischen Lautgeschichte*, Strasbourg: Verlag von Karl J. Trübner.

Meurman-Solin, Anneli (comp.) (1995), *The Helsinki Corpus of Older Scots (HCOS)*, Helsinki: Department of Modern Languages, University of Helsinki.

Stockwell, Robert, and Donka Minkova (1988a), 'The English Vowel Shift: Problems of Coherence and Explanation', in Dieter Kastovsky and Gero Bauer (eds), *Luick Revisited*, Tübingen: Gunter Narr Verlag, 355–394.

The University of Edinburgh (2008–), *A Linguistic Atlas of Older Scots, Phase 1: 1380–1500*, Edinburgh: The University of Edinburgh. Available at: http://www.lel.ed.ac.uk/ihd/laos1/laos1.html.

Stockwell, Robert, and Donka Minkova (1988b), 'A Rejoinder to Lass', in Dieter Kastovsky and Gero Bauer (eds), *Luick Revisited*, Tübingen: Gunter Narr Verlag, 411–417.

Integrating corpus linguistics and NLP methods to explore social media discourse from Edinburgh and London

Lisa Lehnen and Ninja Schulz
(Universität Würzburg)

Cities are highly diverse and dynamic spaces in which multilingual and heterogeneous discourses emerge. The growing body of textual data available from social media (SM) platforms has attracted scholars from various disciplines as a source for real time data generated by "ordinary people" to extract human behaviour, opinions and attitudes and supplement traditional methodologies. By employing natural language processing (NLP) techniques, discourse topics and related evaluations are automatically extracted; however, the potential inaccuracies of automated classifications (see Roberts et al. 2018) are seldom acknowledged. This questions the robustness of the results when viewed through the linguistic lens, e.g. the classification of a tweet as positive or negative tells us little about how people position themselves towards facts and circumstances as various levels of meaning are reduced to polarity. Discourse-analytic (DA) approaches offer tools to unravel these levels of meaning. Moreover, methodological caveats of earlier DA studies regarding reliance on anecdotal evidence and a lack of systematicity have been mended by integrating data-driven procedures (Cheng & Lam 2020: 340): With

corpus-assisted discourse analysis (CADA), "measurable and interpretable linguistic distributions and frequency data lead to the clearer identification of patterns and tendencies within general and specific discourses" (Friginal & Hardy 2020: 1). However, to handle the sheer amount of SM data necessary for the identification of relevant discourse topics in such dynamic spaces as cities, corpus linguistic methods, too, reach their limits. To advance CADA, we integrate data-driven NLP approaches for structuring heterogeneous SM data into meaningful discourse themes, which can serve as a starting point for more fine-grained linguistic analyses.

In this paper, we analyse Twitter discourses during the Covid-19 pandemic, i.e. a time when people's daily routines were disrupted, using geolocated tweets from six administrative areas of both Edinburgh and London, posted between 1 March 2020 and 31 May 2022. These cities represent urban spaces of different sizes, socio-cultural diversity and global connections with London constituting a global megacity, while Edinburgh is a more local cultural centre for Scotland. By running topic modelling, local and global topics were identified, rendering a first indication of differences between the cities. Sentiment analysis provided additional insights into inter- and intra-city variation in the evaluation of these topics. Zooming in on these findings with CADA, we investigate spatio-temporal variation in SM discourses, thus exploring a form of communication with growing importance world-wide which impacts communicative practices. Despite blurring boundaries in the virtual space (Dovchin & Oliver, 2021), we find local colouring in the discourses from the two cities. Thereby this study contributes to evaluate and improve automated tools of analysis and opens new perspectives for corpus-linguistic applications: Assessing how discourses are shaped in urban communities by using SM data is a first step to integrate the complexities of language use and identity constructions in highly heterogeneous urban areas into linguistic theories. Nowadays there is a wealth of (linguistic and nonlinguistic) data available, but we need to develop new methodologies for processing large datasets and find ways to explore and connect them.

References

Cheng, Winnie, and Phoenix Lam (2020), Ideology in media discourse, in S. Conrad, A. J. Hartig, and L. Santelmann (eds), *The Cambridge Introduction to Applied Linguistics*, Cambridge: Cambridge University Press, 339–351.

Dovchin, Sender, and Rhonda Oliver (2021), English and social media: Translingual Englishes, identities and linguascape, in B. Schneider, T. Heyd, and M. Saraceni (eds), *Bloomsbury World Englishes: Paradigms. Volume 1*, London: Bloomsbury Academic, 128–141.

Friginal, Eric, and Jack A. Hardy (2020), Corpus approaches to discourse analysis: Introduction and section overviews, in E. Friginal, and J. A. Hardy (eds), *The Routledge Handbook of Corpus Approaches to Discourse Analysis*, London: Routledge, 1–4.

Roberts, Helen, Bernd Resch, Jon P. Sadler, Lee Chapman, Andreas Petutschnig, and Stefan Zimmer (2018), Investigating the emotional responses of individuals to urban green space using Twitter data: A critical comparison of three different methods of sentiment analysis, *Urban Planning* 3(1), 21–33. Available at: https://doi.org/10.17645/up.v3i1.1231

Variation in clause-final adverbs in a corpus of Colloquial Singapore English

Jakob Leimgruber, JJ Lim, Mie Hiramoto and Wil Gonzales
(University of Regensburg, UC San Diego, National University of Singapore, The Chinese University of Hong Kong)

Clause-final adverbs (CFA) like *already* (1), *also* (2) and *only* (3) are a prominent feature of Colloquial Singapore English (CSE) and other Asian Englishes. While it is also possible for these adverbs to appear clause-finally in standardised varieties of English (e.g., British English), they do so more often in Asian Englishes.

1. Oh okay. Im at fifth floor **alrdy**
'Oh okay. I'm already at the fifth floor.'
<COSEM:18MF02-5714-23INF-2013>
2. I have alot of things to pass to you **also** haha
'I have a lot of things to pass to you too haha.'
<COSEM:17CF34-10659-21CHF-2012>
3. I ate one **only**
'I only ate one.'
<COSEM:18CF55-44567-50CHF-2017>

Such CFA have been extensively studied in works such as Bao and Hong (2006), Cheong (2016), Hiramoto (2015), Parviainen and Fuchs (2019), Teo (2019), Ziegeler (2020), among others. However, little is known about their use in present-day CSE. In this paper Lim et al. 2024) we draw on data from the Corpus of Singapore English Messages (CoSEM; Gonzales et al. 2023) to investigate patterns of variation involving CFA and examine whether they are sensitive to factors such as speaker age and gender.

We follow the argument set forth in Hiramoto (2015) that the clause-finality of *already*, *also*, and *only* is due to the influence of the Sinitic languages spoken in Singapore. Our larger dataset in this paper was collected almost two decades after ICE-SIN, thus allowing to test the stability over time of CFA. We further expected to find instances of sociolinguistic variation with respect to the use of CFA. The hypotheses for this paper are as follows: (i) The use of CFA has increased in CoSEM as compared to ICE-SIN. (ii) Assuming a Sinitic influence, Chinese Singaporeans lead in the use of each adverb in clause-final position. (iii) Inchoative and inceptive functions condition the use of clause-final *already*. (iv) Assuming ongoing change in CSE CFA, certain speakers groups are leading this development, with younger speakers (as leaders in linguistic change, see e.g. Eckert 1989; Labov 2001) using them more frequently.

Our findings suggest that the use of clause-final *already* and *only* have increased over time, while clause-final *also* has remained stable; overall, CFA are found to be a stable feature of CSE. Factors conditioning variation are semantic function, age, and gender. Clause-final *already* is associated with the inceptive function. Clause-final *already* and *also* are more likely to be used by younger speakers, while clause-final *also* and *only* are more likely to be used by males and in all-male conversational settings respectively. We suggest that these patterns are due to present-day English-Mandarin bilingualism, increasingly positive attitudes toward CSE, and National Service for males blurring ethnolinguistic lines in CSE variation.

References

Bao, Zhiming and Huaqing Hong (2006), Diglossia and register variation in Singapore English, *World Englishes* 25(1), 105–114.

Cheong, Phoebe S. E. (2016), Sentence final already and only in Singapore English. Singapore: *National University of Singapore BA (Honours) thesis*.

Eckert, Penelope (1989), The whole woman: Sex and gender differences in variation, *Language Variation and Change* 1(2), 245–267.

Gonzales, Wilkinson Daniel Wong, Mie Hiramoto, Jakob R. E. Leimgruber, and Jun Jie Lim (2023), The Corpus of Singapore English Messages (CoSEM), *World Englishes*, 42(2), 371–388.

Hiramoto, Mie (2015), Sentence-final adverbs in Singapore English and Hong Kong English, *World Englishes*, 34(4), 636–653.

Labov, William (2001), *Principles of Linguistic Change, Volume 2: Social Factors*, Blackwell.

Lim, Jun Jie, Mie Hiramoto, Jakob R. E. Leimgruber, and Wilkinson Daniel Wong Gonzales (2024), Clause-final adverbs in Colloquial Singapore English revisited, *Journal of English Linguistics*, 52(4), 321–345.

Parviainen, Hanna, and Robert Fuchs (2019), I don't get time only: An apparent-time investigation of clause-final focus particles in Asian Englishes, in *Asian Englishes*, 21(3), 285–304.

Teo, Ming Chew (2019), The role of parallel constructions in imposition: A synchronic study of already in Colloquial Singapore English, *Journal of Pidgin and Creole Languages*, 34(2), 346–376.

Ziegeler, Debra (2020), Changes in the functions of already in Singapore English: A grammaticalization approach, *Journal of Pidgin and Creole Languages*, 35(2), 293–331.

Do we write what we feel? Embodiment of emotions in present-day English

Aatu Liimatta, Juha M. Lahnakoski and Ellie Bennett
(University of Helsinki)

The relationship between emotions and the body has been studied in linguistics from various points of view (see e.g. Ogarkova & Soriano 2014: 149). At the same time, psychological research has shown that the ways in which emotions are felt in the body are consistent across cultures (Volynets et al., 2020). However, there is less research on how similar the textual connections between emotions and the body are to the psychological connections between emotions and bodily sensations. That is to say, it is unclear how closely the embodiment of emotions in texts correlates with the way we feel those emotions in our bodies as shown by psychological research.

In this presentation, we aim to answer the following questions:

1. What are the patterns of association between emotion-related and body-related expressions in present-day English; and
2. How do these patterns relate to correlations between emotions and bodily sensations found in experimental psychological research?

Following a methodology developed by Lahnakoski & Bennett et al. (2024), we analyze statistical regularities between the occurrence of emotion-related lexemes (e.g. *anxiety*, *cheerful*, *rejoice*) and body-related lexemes (e.g. *heart*, *thigh*, *tummy*) in the written section of the British National Corpus (BNC). We then combine the emotion-related lexemes into categories corresponding to emotions recognized in psychological research, which enables us to find body-related lexemes which are particularly strongly associated with the different emotion categories.

To measure the strength of association between the emotion and body lexemes, we make use of a simple word embedding model based on the pointwise mutual information of all words in the BNC occurring within a context window of each other (Sahala, n.d.). As opposed to looking for emotion and body lexemes which co-occur in the BNC, the word-embedding approach enables us to find body and emotion lexemes which tend to occur in similar textual contexts and which therefore can be seen as being semantically similar or as representing similar concepts.

We also visualize the results of the analysis using *body maps*. In this visualization technique inspired by psychological research (e.g. Volynets et al., 2020), heatmaps are laid on top of a human silhouette, indicating the strength of association between the body lexemes and the lexemes in the various emotion categories. Finally, we will compare the association patterns and the body maps from the analysis of the BNC with those from psychological research.

Early results indicate that there are clear associations between certain emotion lexemes and body lexemes. For instance, both *love* and *hate* are strongly associated with *heart*, and *nervous* with *stomach*. While the bodily associations of many of the less prototypical emotion words may not be as obvious, we expect based on earlier research that in the aggregate view, patterns between emotions and body-related lemmas will emerge. We also expect to find that the body-emotion associations will largely follow the patterns found in psychological research.

References

Lahnakoski, Juha M., Ellie Bennett, Lauri Nummenmaa, Ulrike Steinert, Mikko Sams, and Saana Svärd (2024), Embodied emotions in ancient Neo-Assyrian texts revealed by bodily mapping of emotional semantics, *iScience* 27(12), 111365. Available at: https://doi.org/10.1016/j.isci.2024.111365

Ogarkova, Anna, and Cristina Soriano (2014), Emotion and the body: A corpus-based investigation of metaphorical containers of anger across languages, *International Journal of Cognitive Linguistics* 5(2), 147–179.

Sahala, Aleksi (n.d.), pmi-embeddings [Computer software]. Available at: https://github.com/asahala/pmi-embeddings

Volynets, Sofia, Enrico Glerean, Jari K. Hietanen, Riitta Hari, and Lauri Nummenmaa (2020), Bodily maps of emotions are culturally universal, *Emotion* 20(7), 1127–1136. Available at: https://doi.org/10.1037/emo0000624

Mind the gap! An experimental appraisal of the task design criteria 'Gap' and 'Meaning Focus'

Donata Lisaitė and Tom Smits
(Kaunas University of Technology, Antwerp Maritime Academy, Antwerp University)

Within the task-based language teaching (TBLT) tradition, the construct *task* is a considerably well-researched area, yet it remains a source of confusion for both researchers and practitioners (e.g., East, 2021; Piccardo C North, 2019; Van den Branden, 2006). We propose a task model based on Ellis and Shintani's criteria for defining a task (2014) grounded in our quasi-experimental study, in which we implemented the task design criteria by using appropriate task instructions to measure the effect of the *Gap* and *Meaning focus* criteria (Ellis C Shintani, 2014) on writing fluency in English as a foreign language. A corpus of writing tasks was created: university students (N= 121) following three study programmes were, per study programme, randomly divided into control and experimental groups and produced five task-based writing samples. Quantitative and qualitative data were collected by using keystroke logging software. The data were first analysed quantitatively in terms of a selection of process- and product-based measures of writing fluency, and on the basis of the results of the quantitative data analysis, a selection of texts was made to analyse them qualitatively (i.e. text quality, perceived fluency). The main findings show that the task design element *Gap* is crucial to fostering writing fluency while the task design criterion *Meaning focus* cannot be confirmed as particularly conducive to fostering writing fluency. These results indicate that including *Gap* as a task design criterion results in a higher level of language users' engagement with the task. On the other hand, the lack of the criterion *Meaning focus* in tasks, i.e. the instruction to focus on language correctness more than on the output, does not result in a substantially higher text quality or considerable differences in terms of the values of the fluency measures among the control and experimental groups.

References

East, Martin (2021), *Foundational Principles of Task-based Language Teaching,* Routledge.

Ellis, Rod and Natsuko Shintani (2014), *Exploring Language Pedagogy through Second Language Acquisition Research,* Routledge.

Piccardo, Enrica and Brian North (2019), *The Action-oriented Approach: a Dynamic Vision of Language Education,* Bristol: Multilingual Matters.

Van den Branden, Kris (2006), *Task-Based Language Education: From Theory to Practice,* Cambridge: Cambridge University Press.

Mapping out types of grammaticalization: A corpus-based quantitative approach

David Lorenz and David Correia Saavedra
(Lund University, Université de Neuchâtel)

A long-standing problem in grammaticalization research has been how to define the boundaries of the concept (i.e. what phenomena are to be subsumed under 'grammaticalization') and what general characteristics can be posited for different types of grammaticalizing items (see von Mengden & Simon 2014). Proposed parameters of grammaticalization (Lehmann 1982, Heine & Kuteva 2007) are empirically useful, but they do not apply consistently to all types of cases that have been considered grammaticalization. Norde & Beijering (2014) propose a clustering approach, in which different cases are sorted by the presence or absence of a set of grammaticalization features to differentiate between grammaticalization, lexicalization and pragmaticalization. Their method highlights important distinctions, yet its applicability remains limited as it relies on outcomes of known changes and thus on prior case studies. Pursuing a similar aim, we present a bottom-up, quantitative approach based on synchronic corpus data.

We use corpus-based metrics of different linguistic elements with the aim of detecting clusters that betray different types of grammaticalization. To this end, we exploit a data set that was created by Correia Saavedra (2021) to quantify the degree of grammaticalization of 528 items in the British National Corpus (BNC Consortium 2007). It contains eight numeric measures associated with grammaticalization: The diversity of collocates (neighboring words, to the left and to the right) and colligates (neighboring word classes, to the left and right) as measures of syntagmatic bondedness and variability; collocate diversity in a wider window to operationalize semantic integrity; word length as a approximation of phonetic integrity / erosion; and token frequency and dispersion, referring to typical usage effects of grammaticalization.

This matrix of eight measures is apt to show similarities and differences between the 528 items. We apply multi-dimensional scaling to map these out in a three-dimensional space. Then, k-means clustering is used to define groups of similar items. Considering the items' positioning and grouping, we explore to what extent they may represent linguistically definable types of grammaticalization. The first results indicate, firstly, a cluster containing mostly lexical nouns and verbs, i.e. non-grammaticalizing items; secondly, at least two clusters of largely grammaticalized items. One of these contains articles, auxiliaries and modal verbs, thus grouping items of high frequency and dispersion and with a high degree of bondedness. Another one seems to groups items that rather show traits of scope expansion and less semantic bleaching (including many complex prepositions). As this work is as yet exploratory, we discuss the analysis and its implications in light of its limitations – such as its dependence on the specific measures – and possible further development, such as a diachronic implementation.

References

Correia Saavedra, David (2021), *Measurements of Grammaticalization: Developing a Quantitative Index for the Study of Grammatical Change,* Berlin: De Gruyter Mouton.

Heine, Bernd and Tania Kuteva (2007), *The Genesis of Grammar: A Reconstruction,* Oxford: Oxford University Press.

Lehmann, Christian (2015 [1982]), *Thoughts on Grammaticalization*, 3rd ed., Berlin: Language Science Press.

Norde, Muriel and Karin Beijering (2014), Facing interfaces: A clustering approach to grammaticalization and related changes, *Folia Linguistica* 48(2), 385–424.

von Mengden, Ferdinand and Horst J. Simon (2014), What is it then, this grammaticalization? *Folia Linguistica* 48(2), 347–360.

The future of corpus linguistics in the age of Large Language Models

Anna Marklová and Jiří Milička
(Charles University)

Work-In-Progress

The fast development and expansion of large language models (LLMs) raises important questions for corpus linguistic. Already today, we have to expect that all newly created corpora contain LLM-generated or assisted texts. This study examines how different these corpora may be from those produced purely by humans, and whether native speakers can reliably distinguish between the two.

Recent advancements indicate that current tools for identifying machine-generated texts exhibit varying levels of accuracy, but overall reliability remains low (Tang, Chuang, & Hu, 2023). In our research, we looked at the characteristics of these texts from the perspective of register variation. We performed multidimensional analysis (MDA) on LLM-generated corpora and discovered that such corpora have reduced register variation compared to human-originated ones.

In the present ongoing study, we focus on specifics of LLM-generated corpora through manual recognition. We address two primary questions: (1) Are people capable of differentiating between human-authored and AI-generated texts, and with what degree of accuracy? (2) Can this ability be developed through feedback? To explore these questions, we present participants with a text drawn from authentic, well-established corpora (Brown Corpus for English, Francis & Kucera 1979; Koditex for Czech, Zasina et. al. 2018) and its corresponding LLM-generated equivalent. The data collection is currently ongoing, the results are expected to be ready by the time of the conference. The participants are asked to guess which one of the texts is AI generated. The speakers are divided into two groups, one group receive immediate feedback if they answer correctly, the other do not. We will analyze if the group with feedback has overall higher scores and if they improve over time. Additionally, we collect information about the attitudes of the speakers towards AI and their confidence in their assessing abilities, and we will measure correlations of these characteristic with the results.

The texts for the experiment are randomly chosen from the corpora, enabling us to control for register effects encoded in the texts´ meta-information. We conduct the experiment on English (Brown Corpus) and on Czech (Koditex corpus), to see how the performance of models differs in the language of the majority of the training data and in a smaller language.

The findings of this study will shed light on the ability of native speakers to recognize machine text production and open a broader debate on the subjective versus objective detectability of AI-generated content. Additionally, we will present quantitative findings (through MDA) and qualitative results (identifying the textual features that assist participants in making their decision) regarding the LLM corpora and how they differ from those comprised solely of human-authored texts. The study will bring an insight into how LLMs change corpus linguistics, and it will open discussion about what tools we have to implement to capture this change in language.

References

Francis, W. Nelson, and Henry Kucera (1979), *Brown Corpus Manual*. Department of Linguistics, Brown University, Providence, Rhode Island, US.

Tang, Ruixiang, Yu-Neng Chuang, and Xia Hu (2023), The science of detecting LLM-generated texts, *arXiv*. https://doi.org/10.48550/arXiv.2303.07205

Zasina, Adrian Jan, David Lukeš, Zuzana Komrsková, Petra Poukarová, and Anna Řehořková (2018), Koditex: korpus diverzifikovaných textů, *Ústav Českého národního korpusu FF UK*.

Explicit and implicit evaluations in comments on the subreddit *Change My View*

Thomas Messerli and Daria Dayter
(University of Basel, Tampere University)

Evaluative practices are central to linguistic persuasion as they intersect with key features of persuasive discourse that have been conceptualised in terms of credibility, stance (DuBois 2007), evaluation (Hunston 2008), appraisal (Martin & White 2005) and positioning (Davies and Harré 2001). For example, studies on liking and persuasion have shown that positive evaluations often serve to establish common ground and bolster relational ties, while negative evaluations, particularly when softened or implicit, can maintain speaker credibility and engagement without overt confrontation (Hyland, 2005; Thompson & Hunston, 2000). Intersubjective alignment (DuBois 2007) similarly is an important tool of creating solidarity and promoting agreement in public discourse, and it is strongly interconnected with implicitness (Harris et al. 2006).

An important distinction in evaluative practices is that between implicit and explicit evaluation (see e.g. Bednarek 2009), i.e. between employing linguistic expressions that explicate a stance towards an object and those that implicate it and are therefore defeasible. Our study focuses on implicit and explicit positive and negative evaluation in the context of persuasion on the subreddit "Change My View" (CMV). Within the digital landscape of typically polarised social media discourse, CMV is a community dedicated to reasoned debate – commenters try to convince original posters (OPs) to change their views. These attempts at persuasion by commenters are governed by an emic marker called *delta*, which is awarded to comments that were seen – typically by the OP themselves – as successfully persuasive.

In our corpus-assisted discourse analysis of two connected corpora containing delta-awarded responses and non-delta-awarded responses, we examine evaluative practices by commenters and hypothesise that while explicit positive and negative evaluation occur in all three corpora, delta-awarded comments will more often employ explicit positive evaluation, a bold on-record positive politeness strategy directed at the original poster, and will typically use more indirect and implicit negative evaluation strategies in order to mitigate potential negative effects on the OP's face. Conversely, we expect non-delta awarded comments to contain a wider range of evaluation strategies that include preferred as well as dispreferred uses of explicitness. We implement the study with a mixed-method design. We start from a wordlist-based approach to explicit evaluation in delta-awarded and non-delta-awarded comments, comparing the relative frequency of positive and negative adjectives extracted from the extant literature across the two corpora. We then select a sample of 100 comments from each corpus based on typicality for the explicit evaluative patterns and manually code them for implicit evaluation. This type of targeted cluster sampling allows us to examine the relative distribution of implicit and explicit evaluative patterns in the sample with a higher likeliness of representativeness for the population.

References

Bednarek, Monika (2009), Dimensions of evaluation: Cognitive and linguistic perspectives, *Pragmatics & Cognition*, *17*(1), 146-175. https://doi.org/10.1075/pc.17.1.05bed

Davies, Bronwyn, and Rom Harré (2001), Positioning the Discursive Production of Selves, in M. Wetherell, S. Taylor, and S. Yates (eds), *Discourse Theory and Practices: A Reader,* Sage, 261–271.

Du Bois, John W. (2007), The stance triangle, in R. Englebretson (ed), *Stancetaking in discourse: subjectivity, evaluation, interaction,* John Benjamins, 139–182.

Harris, Sandra, Karen Grainger, and Louise Mullany (2006), The pragmatics of political apologies, *Discourse & Society* 17, 717–736.

Hunston, Susan (2008), The evaluation of status in multi-modal texts, *Functions of language* 15(1), 64–83.

Hyland, Ken (2005), *Metadiscourse: Exploring Interaction in Writing,* Continuum.

Martin, James R., and Peter R. R. White (2005), *The language of evaluation: Appraisal in English,* Palgrave Macmillan.

Thompson, Geoff and Hunston, Susan (2000), *Evaluation in Text: Authorial Stance and the Construction of Discourse,* Oxford University Press.

Complexity development of intermediate German learners of English: A longitudinal corpus analysis

Philine Kim Metzger
(University of Marburg)

Work-In-Progress

Complexity is a topic of wide range, overall separated into various levels, namely lexis (Linnarud, 1986), morphology (Brezina & Pallotti, 2019), syntax (Lee, 2004) and phraseology (Paquot, 2019). It is generally assumed that syntactical complexity of a language correlates positively with overall language competence and development. (Bulté 2008 & Paquot, Naets, Gries 2021) For example, length of T-Units, clauses per T-Unit or dependent clauses per T- Unit. (Lee 2004: 108) Previous studies already treated the topic, whereas their corpora were much smaller (Kyle, Crossley, Verspoor, 2021). The researchers in the previously mentioned example also worked longitudinal and examined the students throughout years.
The present project therefore poses the following research question:

How does the quantitative and qualitative linguistic complexity in written English of intermediate learners of a German high school develop between 9th and 12th grade?

To answer these research questions, the "Marburg Corpus of Intermediate Learner English" (MILE; Kreyer, 2015) provides the opportunity to analyze data of 90 students from 9th to 12th grade. Within more than 500,000 words in total, the MILE corpus for the first time gives researchers the opportunity for conducting a truly longitudinal analysis of a large number of German intermediate learners of English over four years including different metadata such as gender, age and frequency of usage of English. This corpus will be analyzed using the "Tool for Automatic Analysis of Syntactic Sophistication and Complexity" (TAASSC; Kyle, 2016). Additionally to quantitative aspects, the study conducts teacher's ratings to ensure a qualitative aspect as well.

The method that will be used replicates previous studies connected to syntactic complexity (Crossley & McNamera, 2012) and is called the "bottom-up" principle, which means that data will be analyzed in the first step and will be put into bigger context in the second. The project analyses the complexity development considering Lu's 14 parameters of measuring complexity (Lu 2010: 479).

Results are expected to show the impact of given metadata on complexity development and will be discussed in the light of their language-pedagogical usage. Finally, the project results in an individual and multifactorial learner curve. So far, a pilot study has successfully been conducted and showed how the used tools are appropriate for the specific corpus and first insights on results showed quantitative and qualitative improvement over the course of the years. (Götz-Lehmann, Kettenhofen, Metzger, tba).

References

Brezina, Vaclav, and Gabriele Pallotti (2019), 'Morphological Complexity in Written L2 Texts', *Second Language Research*, 35(1), 99–119.

Bulté, Bram, Alex Housen, Michel Pierrard, and Siska Van Daele (2008), 'Investigating Lexical Proficiency Development Over Time – The Case of Dutch-Speaking Learners of French in Brussels', *Journal of French Language Studies*, 18(3), 277–298.

Crossley, Scott, and Danielle McNamara (2012), 'Predicting Second Language Writing Proficiency: The Roles of Cohesion and Linguistic Sophistication', *Journal of Research in Reading*, 35(2), 115–135.

Götz-Lehmann, Sandra, Fabian Kettenhofen, and Philine Metzger (forthcoming, 2025), *Syntactic Complexity Development of Intermediate L2 English: A Longitudinal, Corpus-Based Study (working title)*.

Kreyer, Rolf (2015), 'The Marburg Corpus of Intermediate Learner English (MILE)', in Marcus Callies and Sandra Götz (eds), *Learner Corpora in Language Testing and Assessment*, Amsterdam: John Benjamins, 13–34.

Kyle, Kristopher (2016), *Measuring Syntactic Development in L2 Writing: Fine-Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication* (Doctoral dissertation, Georgia State University).

Lee, Jiyoung (2004), 'Syntactic Complexity, Clausal Complexity, and Phrasal Complexity in L2 Writing: The Effects of Task Complexity and Task Closure', *The Journal of Asia TEFL*, South Korea, 108–124.

Linnarud, Moira (1986), *Lexis in Composition: A Performance Analysis of Swedish Learners' Written English*, Malmö: C.W.K. Gleerup.

Paquot, Magali (2019), 'The Phraseological Dimension in Interlanguage Complexity Research', *Second Language Research*, 35(1), 121–145.

Paquot, Magali, Hubert Naets, and Stefan Th. Gries (2021), 'Using Syntactic Co-occurrences to Trace Phraseological Complexity Development in Learner Writing: Verb + Object Structures in LONGDALE', in Bert Le Bruyn, Walter Simonne, and Magali Paquot (eds), *Learner Corpus Research Meets Second Language Acquisition*, Cambridge: Cambridge University Press.

Corpus Sense: A next-generation tool for advanced corpus and discourse analysis

Antonio Moreno-Ortiz
(University of Malaga)

Software demonstration

Corpus Sense is a corpus query tool that incorporates advanced functionalities not available in existing applications. It is specially designed for content and discourse analysis, although it also features functionalities commonly found in other corpus tools. Corpus Sense combines quantitative, qualitative and AI features to offer users a unique set of tools that is able to easily obtain useful insights of a corpus with minimal effort.

Corpus Sense is a web application written in Python that is designed to work with small to medium-sized corpora (up to 2 million tokens). As such, it offers user registration and management capabilities, including corpus upload and sharing (public corpora). It offers support for 22 languages, although no support is given to multilingual corpora.

Unlike other corpus tools, Corpus Sense uses a Natural Language Processing framework, Spacy (Honnibal et al., 2020), to process and query corpora, and also creates word embeddings using Transformers (Wolf et al., 2020), which are used for several of the content analysis features. This means that no actual Database Management System is actually used to store the corpus, all corpus data is stored as Pandas dataframes and binary objects, which are serialized after processing and later loaded on demand. Thus, the application is, in general, extremely responsive for common operations, such as searching, and offers advanced, NLP-based features, such as graph-based keyword extraction (thus eliminating the need for a reference corpus), named entity recognition and labeling, and literal, pattern-based, and semantic search.

Additionally, state-of-the-art topic modeling is offered by means of a user-friendly interface to BERTopic (Grootendorst, 2022), a sophisticated topic modeling library. Another defining, advanced feature, tentatively named "Insights", leverages the power of large language models to produce high-quality descriptions of variety of aspects of the contents of a corpus. These aspects include rhetorical style and devices, discourse markers, readability, sentiment, emotions, hate speech detection, etc. Unlike most other AI tools, Corpus Sense does not make use of external LLMs through an API; instead, it runs a freely available multilingual LLM locally, currently Qwen2.5 (Qwen Team, 2024/2024). Because of the multilingual capabilities of this LLM, insights can be generated in many languages, regardless of the corpus language.

The following summarizes the different pages/tools of the application:

- The *Snapshot* offers an LLM-generated overview of the contents of the corpus. It also contains corpus statistics, keywords (single-word and multi-word), labeled named entities, and frequency lists of common Internet symbols (emojis, hashtags, and mentions).
- *Words*: token and lemma lists are easily filtered by a number of criteria, including part-of-speech tags and substrings.
- *Search*: semantic search based on word embeddings is built-in, as are lexical search (with regular expressions support) and token-based pattern search.
- *Topics*: identified topics are actually given readable descriptive labels (generated by the local LLM) instead of the typical set of keywords. The tool allows users to view relevant text snippets for each identified topic.
- *Insights* are created using a user-friendly wizard, which allows selection of the data source as well as the desired aspect. Identified entities, topics and even the results of a semantic search can be used as data sources.
- *Subcorpora* allows users to create one or more subcorpora based on the files of a corpus. Once a subcorpus is created and activated, all tools will show options and results for that specific subcorpus.

Corpus Sense is still under development and new users can only register via pre-authorization. A pre-authorized account can be provided on request using the contact form of the app.

References

Grootendorst, Maarten (2022), *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. https://doi.org/10.48550/ARXIV.2203.05794

Montani, Ines, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters (2020), *spaCy: Industrial-strength Natural Language Processing in Python* [Computer software], Zenodo. https://doi.org/10.5281/zenodo.1212303

Qwen Team (2024), *Qwen2.5: A Party of Foundation Models* [Shell], Qwen. https://github.com/QwenLM/Qwen2.5 (Original work published 2024)

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (2020), Transformers: State-of-the-art natural language processing, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

A new resource for the study of late medieval English: The Armburgh Correspondence

Terttu Nevalainen and Sara Norja
(University of Helsinki)

Personal letters provide a window on individual language use in the past. They are amply available from the more recent periods but become scarcer as we go further back in time. For example, the 18th-century extension of the *Corpus of Early English Correspondence* (CEEC) includes material from 77 letter collections but the 15th-century section from half a dozen, with only four major ones: the Cely, Paston, Plumpton and Stonor collections. The publication in 1998 of the Armburgh papers adds one more source to this small number of published collections of late medieval English personal correspondence.

After introducing the *Armburgh Correspondence Corpus* (ACC), based on the published Armburgh Papers, this presentation will discuss its relevance as material for historical sociolinguistic and pragmatic research. Case studies are provided to show how the language of the ACC relates to findings on other comparable contemporary resources and its potential future uses.

The letters included in the ACC were written to and by the members of the Armburgh family of Mancetter, Warwickshire, between the 1420s and 1450s. As was the case with family letter collections such as the Paston letters, they were occasioned by a major lawsuit. Much of the correspondence was by one of the claimants, Joan Armburgh (née Brokholes), and her husband Robert, concerning the Brokholes inheritance scattered over Warwickshire, Huntingdonshire, Hertfordshire and Essex. The original manuscript contains transcripts of c. 100 letters and memoranda, but the ACC currently includes only the 70 letters related to this land dispute. All of them have been preserved as copies, produced by four scribes. Although the scope of the ACC is more limited than that of the major 15th-century letter collections, and the Paston letters in particular, it provides unique evidence on individual and group usage and on the family and feudal networks of the time. The linguistic case studies to be discussed in the paper include the choice of second-person singular pronouns (*you* vs. *thou*), forms of address, and verb morphology (3rd person sg. and pl. present tense indicative forms, infinitive forms, and *be* vs. *are* variation). The results show both grammatical consistency and contextual register variation.

References

Carpenter, Christine (1998), *The Armburgh Papers: The Brokholes Inheritance in Warwickshire, Hertfordshire, and Essex, c. 1417– c. 1453: Chetham's Manuscript Mun. E. 6.10(4),* Rochester, NY: Boydell & Brewer.

Nevalainen, Terttu, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Jukka Keränen, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily, and Anni Sairio (1998-2006), *Corpus of Early English Correspondence,* Department of Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC/

Assessment of a phraseographical methodology and model developed for a corpus-based multilingual platform of Collocations Dictionary

Adriane Orenha-Ottaiano and Maria Eugênia Olímpio de Oliveira Silva
(São Paulo State University, University of Alcalá)

In this paper, we present and discuss a phraseographical methodology and model that has been developed for the phraseographical treatment of a set of collocations that make up a lexicographical resource (under construction). The proposal is part of the project *A Phraseographical Methodology and Model for an Online Corpus-Based Multilingual Collocations Dictionary Platform* (FAPESP Process no. 2020/01783-2), whose main objective is to develop a *Multilingual Platform of Collocations Dictionary* (PLATCOL), in English, French, Portuguese, Spanish, and Chinese – Italian, German, and European Portuguese in the second phase of the project. It will operate in an open-access environment and will allow online consultation of the lexical units referred to. It is a resource aimed at language learners, teachers, creators of teaching materials as well as professional and student translators (Orenha-Ottaiano et al., 2021). The methodology integrates automatic extraction techniques, the use of PLN tools to annotate corpora with lemmas, grammatical categories and dependency relations. Subsequently, statistical measures and distributional semantics strategies were applied to select candidates for collocations (Evert et al., 2017; Garcia et al., 2019). To build the platform, large corpora were compiled and processed with NLP tools: *UDPipe* (English, French), *LinguaKit* (Portuguese, Spanish), and *Stanford CoreNLP* (Chinese). Tokenization and morphosyntactic tagging used state-of-the-art models for accurate annotation (Straka & Straková, 2017). Researchers have then carefully post-edited the results for quality. So far, the following achievements have been made in the framework of this project: a) the construction of a database of collocations (with a total of 203,051 registered candidates: 51,574 in English; 60,215 in French; 63,615 in Portuguese; 427 in Mandarin and 27,220 in Spanish), and b) the development of a software, the *Collocations Dictionary Writing System* (COLDWS), which allows researchers to access the database and work simultaneously on the analysis of these data. PLATCOL's phraseographic proposal is based on the contributions of the Theory of Lexicographic Functions (Fuertes

Olivera & Tarp, 2014; Tarp, 2014, 2015, etc.) and, consequently, pursues the purpose of meeting the needs of the platform's potential users and fostering their collocational knowledge. For this reason, in the development of the phraseographic methodology and model presented here, it has been essential, on the one hand, to innovate in the phraseographic description of these lexical units, going beyond the indication of a lemma and its syntactic classification, and the assignment of a semantic label. On the other hand, it has also been necessary to pay special attention to the way in which this phraseographic information should be presented to users. The aim is to facilitate access to the data and thus the understanding of the data presented. In line with the above, the phraseographic model being implemented will incorporate information about the syntagmatic combinations of the units, explanations of semantic nature, examples and usage notes (frequent constructions, collocations and/or related combinations, etc.). In turn, the methodology developed adopts a corpus-based approach, so that the analysis of the data recorded in COLDWS is complemented by the review of data extracted from dictionaries and other corpora.

References

Evert, Stefan, Peter Uhrig, Sabine Bartsch, and Thomas Proisl (2017), E-VIEW-affilation – A large-scale evaluation study of association measures for collocation identification, in I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, and V. Baisa (eds), *Proceedings of eLex 2017–Electronic lexicography in the 21st century*: *Lexicography from Scratch,* Leiden, the Netherlands, 531–549. Available at: https://elex.link/elex2017/proceedings/eLex_2017_Proceedings.pdf.

Fuertes-Olivera, Pedro A. and Sven Tarp (2014), *Theory and Practice of Specialised Dictionaries, Lexicography versus Terminography*, Berlín/Boston: Walter de Gruyter.

Garcia, Marcos, Marcos García-Salido, and Margarita Alonso-Ramos (2019), A comparison of statistical association measures for identifying dependency-based collocations in various languages, in A. Savary, C. Parra Escartín, F. Bond, J. Mitrović, and V. B. Mititelu (eds), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet* (MWE-WN 2019), Florence, Italy, 49-59. Available at: https://www.aclweb.org/anthology/W19-5107.pdf.

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychly (2008), GDEX: Automatically Finding Good Dictionary Examples in a Corpus, in E. Bernal and J. DeCesaris (eds), *Proceedings of the 13th EURALEX International Congress*. Barcelona: Institut Universitari de Linguistica Aplicada/Universitat Pompeu Fabra, 425–432.

Orenha-Ottaiano, Adriane, Marcos Garcia-Gonzalez, Maria Eugênia Olímpio de Oliveira Silva, Marie-Claude L'Homme, Margarita Alonso Ramos, Carlos Roberto Valencio, and William Tenorio (2021), Corpus-based Methodology for an Online Multilingual Collocations Dictionary: First Steps, in *Proceedings of Elex2021*, 1–28.

Straka, Milan and Jana Straková (2017), Tokenizing, POS-tagging, lemmatizing and parsing UD 2.0 with UDPipe, in J. Hajič and D. Zeman (eds), *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies,* Vancouver, Canada, 88-99. Available at: https://www.aclweb.org/anthology/K17-3009.pdf.

Tarp, Sven (2014), Theory-Based Lexicographical Methods in a Functional Perspective. An Overview [Theoriebasierte lexikographische Methoden aus funktionaler Perspektive. Ein Überblick / Méthodes en lexicographie théorique du point de vue fonctionnel. Une vue d'ensemble], *Lexicographica* 30(1), 58–76.

Tarp, Sven (2015), La teoría funcional en pocas palabras, E*studios de Lexicografía* 4, 31–42.

Development of grammatical text complexity in learner English at university level

Jane Padrik, Denys Savchenko and Janely Rüdein
(University of Tartu)

The study focuses on the use of grammatical complexity features in university level texts written by learners of English at both undergraduate and graduate levels. Corpus-based studies have suggested that there are two fundamentally different types of grammatical complexity: phrasal and clausal (Biber 1992; Biber & Gray 2010, 2011, 2016; Biber et al. 2011, 2024; Staples et al. 2016). Previous studies have shown two trends as academic level of students increases: 1) the use of phrasal complexity features in writing increases; 2) clausal complexity features in student writing, particularly finite dependent clauses, decreases. It has been shown (Lu 2011; Parkinson & Musgrave 2014; Biber et al. 2014) that student writers vary in their use of phrasal complexity features across level. However, those differences cannot be fully understood without accounting for their L1 background. While the primary focus of the analysis is on writing development from third-year undergraduate to graduate students by Estonian learners of English, we also look at L1 English writers and Swedish learners of English.

We address the following research questions:

1. Is the general hypothesis that L2 English learners develop in the use of grammatical complexity across the university years supported by corpus evidence?
2. How are the patterns of development mediated by learners' L1?

For this purpose, we examine the development of phrasal and clausal complexity features in L2 English learners at two levels of study (final-year undergraduate and graduate level) from two L1 backgrounds (Estonian and Swedish learners of English). We base our analysis on three corpora: BAWE (Gardner & Nesi 2013), VESPA (Paquot et al. 2022) and TCELE (the Tartu Corpus of Estonian Learner English). The Estonian learner data comprises 75 BA (491,198 words) and 21 MA theses (712,946 words), with MA theses selected to include texts by the same learners as in the BA component to track individual development. The Swedish learner data comprises 51 theses (364,174 words) from VESPA. To ensure comparability, the BAWE corpus was sampled to be as similar as possible to the learner data with regard to text type and discipline. In total, 295 texts were sampled from BAWE (692,683 words).

The grammatical complexity analysis was based on roughly 20 grammatical features that have been identified in previous research (Biber et al. 2014, 2020, 2024; Staples et al. 2016). These include clausal structures (finite and non-finite dependent clauses) and phrasal structures. For the pilot study, we automatically annotated the corpus texts using the Multidimensional Analysis Tagger (MAT; Nini 2019). To investigate the differences in grammatical complexity measures across level of study and learners' L1, we fitted different mixed-effects regression models using the statistical environment R. The preliminary results indicate that clausal and phrasal complexity features differed systematically across levels of study for Estonian learners of English – phrasal features increase across level of study, while clausal features decrease. The study adds to the previous findings by more systematically showing the development across a wide set of linguistic features by comparing L2 writing from two L1 backgrounds.

References

Biber, Douglas (1992), On the complexity of discourse complexity: A multidimensional analysis, *Discourse Processes* 15, 133–163.

Biber, Douglas, and Bethany Gray (2010), Challenging stereotypes about academic writing: Complexity, elaboration, explicitness, *Journal of English for Academic Purposes* 9, 2–20.

Biber, Douglas, and Bethany Gray (2011), The historical shift of scientific academic prose in English towards less explicit styles of expression: Writing without verbs, in V. Bhatia, P. Sánchez, and P. Pérez-Paredes (eds), *Researching Specialized Languages*, Amsterdam: John Benjamins, 11–24.

Biber, Douglas, and Bethany Gray (2016), *Grammatical Complexity in Academic English: Linguistic Change in Writing*, Cambridge: Cambridge University Press.

Biber, Douglas, Bethany Gray, and Kornwipa Poonpon (2011), Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45, 5–35.

Biber, Douglas, Bethany Gray, and Shelley Staples (2014), Predicting patterns of grammatical complexity across language exam task types and proficiency levels, *Applied Linguistics*, Advance online publication.

Biber, Douglas, Tove Larsson, and Gregory R. Hancock (2024), The linguistic organization of grammatical text complexity: Comparing the empirical adequacy of theory-based models, *Corpus Linguistics and Linguistic Theory* 20(2), 347–373.

Biber, Douglas, Bethany Gray, Shelley Staples, and Jesse Egbert (2020), Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement, *Journal of English for Academic Purposes* 46, 100869.

Gardner, Sheena, and Hilary Nesi (2013), A classification of genre families in university student writing, *Applied Linguistics* 34(1), 1–29.

Lu, Xiaofei (2011), A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writer's language development, *TESOL Quarterly* 45, 36–61.

Nini, Andrea (2019), The Multi-Dimensional Analysis Tagger, in T. Berber Sardinha, and M. Veirano Pinto (eds), *Multi-Dimensional Analysis: Research Methods and Current Issues*, London and New York: Bloomsbury Academic, 67–94.

Paquot, Magali, Tove Larsson, Hilde Hasselgård, Signe Oksefjell Ebeling, Damien De Meyere, Larry Valentin, Natalia J. Laso, Isabel Verdaguer, and Sanne van Vuuren (2022), The Varieties of English for Specific Purposes database (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing, *Research in Corpus Linguistics* 10(2), 1–15.

Parkinson, Jean, and Jill Musgrave (2014), Development of noun phrase complexity in the writing of English for academic purposes students, *Journal of English for Academic Purposes* 14, 48–59.

Staples, Shelley, Jesse Egbert, Douglas Biber, and Bethany Gray (2016), Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre, *Written Communication* 33(2), 149–183.

Explanatory processes in *Ask an Expert* websites on psychology and mental health

Daniel Pascual
(University of Zaragoza)

Psychological issues and mental health have gained tremendous relevance in daily life and, consequently, in public discourse and digital interactions. Users frequently seek advice and trustworthy information to satisfy their curiosity and address pressing concerns. However, the growing prevalence of online misinformation and harmful practices poses challenges for users searching for reliable resources. Within this context, *Ask an Expert* practices offer valuable spaces where qualified professionals, validated for their expertise and academic credentials, respond to user queries. *Ask an Expert* practices have been examined in specific domains like healthcare (Pounds 2018) and environmental sustainability (Pascual 2025), there is a need to explore how specialised knowledge is adapted and presented in these practices when distributing information on psychological and mental health issues to a diverse audience of non-experts users. This research also complements recent studies on other digital practices geared towards dissemination in psychology, like research digests (Lorés 2024).

This study aims, first, to identify the use and frequency of verbal explanatory processes employed by professionals on *Ask an Expert* websites to recontextualise psychological and mental health information. Second, it examines how these processes are realised through specific discursive mechanisms. To that end, a corpus of 100 texts retrieved from four prominent sources (*GoAskAlice!*, *IDONTMIND*, *PsychHelp* and *Child Mind Institute*) has been compiled from the SciDis Database, which is a collection of digital practices exemplifying contemporary scientific dissemination (Pascual and Sancho-Ortiz 2024). The four websites

selected are characterised by including a devoted section in which professionals working for universities and accredited associations take up users-submitted questions on psychological concerns. The analysis rests on Mur-Dueñas' (2024) taxonomy of explanatory processes for the field of economy, including exemplification, explicitation, elaboration, enumeration, and comparison/analogy. Using the qualitative analysis software NVivo12, the study identifies how these processes work in the field of psychology, and examines their frequency and implementation in general terms as well as across the four websites chosen. Findings show connections between such explanatory processes and specific discursive mechanisms (lexico-grammatical, syntactic, orthotypographic), and illustrate how experts support the simplification and easification of presumably complex concepts when fostering knowledge dissemination to non-expert audiences. Overall, this paper sheds light on experts' assumptions regarding users' information needs and, more interestingly, offers guidance for professionals to leverage discursive strategies and mechanisms, which they can use to bridge knowledge gaps and make psychological information more accessible and impactful for diversified digital audiences.

References

Lorés, Rosa (2024), Digesting psychology: Metadiscourse as a recontextualizing tool in the digital communication of disciplinary research, *Chinese Journal of Applied Linguistics* 47(2), 178–195. https://doi.org/10.1515/CJAL-2024-0202

Mur-Dueñas, Pilar (2024), Digital dissemination practices: An analysis of explanatory strategies in the process of recontextualising specialised knowledge, *Discourse & Interaction* 17(1), 94–114. https://doi.org/10.5817/DI2024-1-94

Pascual, Daniel (2025), Dialogic markers in *Ask an Expert* webpages on environmental discourse, *Language & Dialogue* 15(1), 156–181. https://doi.org/10.1075/ld.00191.pas

Pascual, Daniel and Ana Eugenia Sancho-Ortiz (2024), Investigating recontextualisation processes in scientific digital practices: The SciDis database, *Revista Electrónica de Lingüística Aplicada* 23(1), 101–118.

Pounds, Gabrina (2018), Patient-centred communication in Ask-the-Expert healthcare websites, *Applied Linguistics* 39(2), 117–134. https://doi.org/10.1093/applin/amv073

Discourse-pragmatic *LIKE* in computer-mediated written discourse

Veronika Raušová
(Charles University)

Discourse-pragmatic (D-P) *LIKE* serves as a versatile interpretive cue employed in casual spoken conversation to navigate the structure of discourse and guide the decoding of communicative intent. Due to its ubiquity, it has been extensively investigated (e.g., Andersen, 2001; Schweinberger, 2014; D'Arcy, 2017). Despite this, its frequency in casual conversations warrants periodic re-evaluation, especially as everyday interactions increasingly shift online.

The present research builds on Raušová's (2023) study, which compared the frequency and functions of D-P *LIKE* between the demographically sampled spoken component of the BNC1994 and the Spoken BNC2014 (Love et al., 2017) and was motivated by the significant increase in relative frequency of the word *LIKE* in the latter corpus (from 4,370.6 i.p.m. to 13,781.9 i.p.m.). Her findings attribute the growth to its D-P functions, including *LIKE* as a discourse marker signalling relationships between discourse segments (1) or accompanying disfluencies (2); a pragmatic marker in clause-medial (3) or clause-final (4) positions highlighting sentence elements or suggesting a need for non-literal interpretation; and *LIKE* as part of quotative frames (5) introducing speech, thoughts, or re-enactments (Raušová, 2023).

1. S0439: *I am very British in my emails I think* ***like*** *I'm very not formal* […]
2. S0520: *you know some some* ***like*** *all of us have our things* […]

(3) S0529: *although our cinema's* ***like*** *really tiny* […]
(4) PS1JP: […] *she's killing herself for the job* ***like****!*
(5) S0235: […] ***he was like*** *(.) no he's Romanian* ***I was like*** *no he's not* […] (Raušová, 2023)

This study investigates whether D-P *LIKE* extends its functionality to computer-mediated written discourse (CMWD), granted that CMWD (e.g., online forum comments) exhibits features of casual spoken discourse (Cutler et al., 2022). It is hypothesized that most of the D-P functions described in Raušová (2023) will be attested, except for *LIKE* as a discourse marker accompanying disfluencies, which seems to be incompatible with written communication's asynchronous, editable nature. The study draws on a working corpus of Reddit posts and associated comments compiled using the PRAW Python package (Boe, n.d.). Preliminary findings are based on a randomly sampled dataset of 1,000 occurrences of the word *LIKE*, which were qualitatively analysed following the methodology outlined by Raušová (2023).

The findings confirm the presence of D-P *LIKE* in CMWD, with most functions from Raušová's (2023) taxonomy attested. Notably, *LIKE* most frequently appeared as a discourse marker with discourse linking function. As expected, *LIKE* as a disfluency marker was absent in the examined data, along with the clause-final pragmatic marker. The absence of the latter can be explained by its regional ties to dialects of northern England, Scottish English, and Irish English (Schweinberger, 2014: 354), while Reddit's user base is predominantly located in the U.S.

The study aims to demonstrate the intentionality of D-P *LIKE*, further countering interpretations of its use as merely symptomatic of planning difficulties. By examining its transition from in-person spoken to online written discourse, the findings provide insights into how discourse-pragmatic features transfer and adapt across communication modes.

References

Andersen, Gisle (2001), *Pragmatic Markers and Sociolinguistic Variation: A relevance-theoretic approach to the language of adolescents*, Amsterdam/Philadelphia: John Benjamins Publishing Company. [Pragmatics & Beyond New Series, 84], ix, 352 pp. https://doi.org/10.1075/pbns.84

Boe, Bryce (n.d.), *PRAW: The Python Reddit API Wrapper* (Version 7.8.1) [Computer software], PRAW. https://praw.readthedocs.io/

Cutler, Cecelia, May Ahmar, and Soubeika Bahri (2022), Introduction: The oralization of digital written communication, in C. Cutler, M. Ahmar, and S. Bahri (eds), *Digital Orality: Vernacular Writing in Online Spaces*, Cham: Springer International Publishing, 3–31.

D'Arcy, Alexandra (ed) (2017), *Discourse-Pragmatic Variation in Context: Eight hundred years of LIKE*, Amsterdam/Philadelphia: John Benjamins Publishing Company. [Studies in Language Companion Series, 187], xx, 235 pp. https://doi.org/10.1075/slcs.187

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery (2017), The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations, *International Journal of Corpus Linguistics* 22(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Raušová, Veronika (2023), *Discourse-Pragmatic Functions of Like in Spoken Discourse* [Doctoral dissertation, Charles University, Faculty of Arts], Charles University Digital Repository. http://hdl.handle.net/20.500.11956/188271

Schweinberger, Martin (2014), *The Discourse Marker LIKE: A Corpus-Based Analysis of Selected Varieties of English* [Doctoral dissertation], The University of Queensland. https://doi.org/10.13140/RG.2.2.28150.65603

Tracking change in recent American English: The case of BE *saying*

Paula Rautionaho
(University of Eastern Finland)

The co-occurrence of communication verbs, such as *talk* or *say*, with the progressive form has increased tremendously in recent years (see e.g. Rautionaho, in press). Rohe (2019), for instance, ties this development to the increase of non-aspectual uses of the progressive, such as recentness or interpretative progressives (see (1) and (2), respectively).

1. perhaps this has something to do with what **I was just saying** about a lack of education (BNC2014, WtsDra274)
2. relax, I'm not accusing you of anything! **I'm just saying** you like her (BNC2014, FictMis121)

This study focuses on recent American English in an effort to map out the different forms and functions of the progressive BE *saying* in the years 1990–2019. Having extracted all instances of BE *saying* in the *Corpus of Contemporary American English* (COCA, N=62,113), I have identified different patterns with *saying*, such as *as X* BE *saying*, *what X* BE *saying*, and *X* BE *just saying*, and will discuss how their frequency has evolved in the three decades under scrutiny. The research questions are as follows:

i. To what extent do the different patterns with BE *saying* share similarities in their development in COCA?
ii. To what extent does the increase of pragmatic uses of the progressive explain the increasing use of BE *saying* overall?

Methodologically, this study employs Variability-based Neighbor Clustering (VNC; Gries & Hilpert 2012) to partition the data into meaningful time periods, after which the evolutionary patterns of different realisations of BE *saying* are investigated in detail. Further, Generalized Linear Mixed Model (GLMM) Tree analyses (Fokkema et al. 2020) are modelled to assess the 'progressivity' of the identified patterns, that is, to assess to what extent the frequency developments are tied to the progressive form as opposed to the non-progressive counterparts (as in *what I said*, *I just said*). Finally, the findings of the study are discussed in the light of changes in the progressive paradigm overall. While Rohe (2019) reports on increasing use of meta-communicative and pragmatic uses of the progressive in recent British English, the results here address similar patterns in recent American English.

Preliminary results of the study confirm the increasing use of BE *saying* also in AmE and indicate that the identified patterns show deviant evolutionary patterns, with *X* BE *just saying* showing more pronounced increase especially in the latest period investigated. The GLMM Trees indicate that the patterns are highly 'progressivised', which lends support to the evolving nature of the progressive form overall.

References

BNC2014 (2014), *The British National Corpus 2014*. Available at: http://corpora.lancs.ac.uk/bnc2014/

Davies, Mark (2008-), *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

Gries, Stefan Thomas, and Martin Hilpert (2012), Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics, in T. Nevalainen, and E. C. Traugott (eds), *The Oxford Handbook of the History of English*, Oxford: Oxford University Press, 134–144.

Fokkema, Marjolein, Julian Edbrooke-Childs, and Miranda Wolpert (2020), Generalized linear mixed-model (glmm) trees: A flexible decision-tree method for multilevel and longitudinal data, *Psychotherapy Research* 31(22), 1–13.

Rautionaho, Paula (in press), Grinding to a halt? The spread of the progressive in recent spoken British English, in A. Carlucci, and J. Nykiel (eds), *The Progressive Revisited: Studies on Germanic and Romance Languages*, *Studies in Language Companion Series*, Amsterdam: John Benjamins.

Rohe, Udo Julius (2019), *The Progressive in Present-Day Spoken English: Real-Time Studies of Its Spread and Functional Diversification*, PhD dissertation, Albert-Ludwigs-Universität. Available at: https://d-nb.info/1196526699/34

A cluster-based linguistic analysis of propaganda techniques in Lithuanian news articles

Ieva Rizgelienė, Evelina Vaitkevičiūtė, Gražina Korvel and Vilma Zubaitienė
(Vilnius University)

Work-In-Progress

Propaganda is a phenomenon that has been extensively studied in academic research. It can be defined as the dissemination of information, ideas or rumours with the intention of helping or harming a particular institution, person or group. The methods used in propaganda are diverse, but the literature consistently points to several prevalent techniques. These include loaded language, emotional appeals, and selective, biased, or misleading information. This study focuses on using machine learning methods to perform a cluster-based linguistic analysis of Lithuanian news articles, with the aim of identifying the keywords and syntactic structures that characterize the different propaganda techniques.

The dataset used in this study consists of Lithuanian news articles annotated by linguists and social scientists as part of the project on propaganda and disinformation research (ATSPARA). This collection focuses on 10 propaganda techniques annotated through a cross-annotation process. The dataset is still in the process of compilation during the preparation of this research, so it is a preliminary version of the entire dataset.

The research began with a statistical analysis to determine the distribution of each propaganda technique within the dataset. Following the distributional analysis, the study delved into a linguistic exploration of each propaganda technique. Natural language processing techniques such as tokenization, part-of-speech tagging, and syntactic parsing were used to extract and examine in detail key linguistic features, facilitating the identification of specific words and syntactic structures commonly associated with each propaganda technique. In addition, dependency parsing was used to uncover deeper syntactic relationships and structures within the sentences, providing insight into the use of complex linguistic patterns.

To validate the accuracy of the extracted linguistic information, KMeans and hierarchical clustering algorithms were applied. This approach allowed us to conduct a detailed examination of how different propaganda techniques cluster together based on extracted linguistic features, as well as how they diverge in their use of language. The clustering results provided a nuanced understanding of the relationships and differences between the techniques, highlighting patterns that were not immediately evident through manual analysis.

The findings of this research underscore the complex use of language in propaganda, highlighting both common and unique aspects of different techniques. These results contribute to the field of linguistic analysis by providing a detailed framework for understanding and identifying propaganda in news articles, offering practical applications for detecting manipulative content in media.

Analysing Donald Trump's political rhetoric

Patricia Ronan and Gerold Schneider
(University of Dortmund, University of Zurich)

Donald Trump, the 45th President of USA has been re-elected President, accompanied by discussions about populism and right-wing tendencies and rhetoric. That President Trump has shown signs of populist language has been discussed by various authors (e.g., Oliver & Rahn 2016, McDonnell & Ondelli 2022). Arguably, various features of the language of right-wing populism (Wodak 2015) can also be found.

The current study aims to find out 1) whether right-wing populist features can be found in Donald Trump's political rhetoric, 2) whether use and intensity of right-wing populist features have been changing over time.

In order to answer these research questions, features of right wing populism determined by Wodak (2015) are traced in transcriptions of political interviews with Donald Trump. The interviews are taken from the Roll Call website at, rollcall.com/factbase, which offers transcripts of checkable soundfiles. From this source, a corpus of political interviews of approximately 50,000 words per decade (1980s -2020s) is created and analysed qualitatively for the presence of right-wing tropes. We first use manual annotation of a large sample and, second, compare it to an automatic approach using GPT (OpenAI et al. 2023) or a similar large language model, for example Llama3 (Grattafiori et al. 2024). Second, in order to evaluate, we compare the results of the automatic and the manual approach. In the third step, we apply the automatic approach to the entire corpus, thus allowing us to perform a qualitiative analysis at large scale, and also enabling a robust quantitative analysis. After the presence and the nature of right-wing tropes is determined, a sentiment analysis is carried out with SentimentR to determine the polarity of the expressed sentiments.

Results show which right-wing tropes can be found in Donald Trump's political language and how their focus has been shifting during the last three decades.

References

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, and others (2024), The Llama 3 herd of models, *arXiv:2407.21783*. Available at: https://doi.org/10.48550/arXiv.2407.21783

McDonnell, Duncan, and Stefano Ondelli (2022), The language of right-wing populist leaders: Not so simple, *Perspectives on Politics* 20(3), 828–841.

Oliver, J. Eric, and Wendy M. Rahn (2016), Rise of the Trumpenvolk: Populism in the 2016 election, *Annals of the American Academy of Political and Social Science* 667, 89–206.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and others (2024), GPT-4 technical report, *arXiv preprint arXiv:2303.08774*. Available at: https://doi.org/10.48550/arXiv.2303.08774

Wodak, Ruth (2015), *The Politics of Fear: What Right-Wing Populist Discourses Mean*, London: Sage.

AI vs. human writing: A corpus-based comparative study of lexical and grammatical features

Karolina Rudnicka
(University of Gdansk)

The advent of AI-powered technologies, such as natural language generation models, has sparked global debate. While some researchers see it as a transformative revolution offering vast opportunities (Harari, 2017; Aubignat & Diab, 2023), others warn of potential risks (De Angelis et al., 2023; Gul, 2023). Still, others view these concerns as exaggerated (e.g. Leaver & Srdarov, 2023). Beyond these debates, AI-generated language output has become an intriguing subject for linguists, particularly given the dominance of English in global communication among native and non-native speakers alike.

Existing research has predominantly focused on AI's applications in English language teaching and learning (O'Neill & Russel, 2019; Barrot, 2020). However, the linguistic impact of AI on textual output and the English language itself remains underexplored. Addressing this gap, this study examines the variation and diversity in written English as produced by humans and AI, focusing on how grammatical and lexical features differ between the two. The study builds on a prior pilot investigation (Rudnicka, 2023) that explored the effects of tools like Grammarly and ChatGPT on conciseness versus wordiness. The present

work uses a subset of the Human-ChatGPT (gpt4-o) comparison corpus[1], comprising human-written TOEFL essays and ChatGPT-generated responses to the same questions. The essays are analysed with focus on lexical diversity (Type-Token Ratio - TTR) and grammatical variation. R, along with the 'tm' package, is employed to clean the data and manipulate the texts.

Preliminary analysis of five essays per group shows notable patterns. ChatGPT- generated essays exhibit consistent and moderate lexical diversity (TTR: 0.48–0.54), suggesting a systematic approach balancing vocabulary variety and repetition. In contrast, human-written essays show greater variability (TTR: 0.33–0.51), reflecting individualized writing styles and topic familiarity. Additionally, ChatGPT essays are more verbose, with a larger vocabulary.

This study situates itself between small-scale analyses (Skowron & Bączkowska, 2023) and large-scale comparisons (Herbold et al., 2023; Reviriego et al., 2024), contributing a qualitative perspective alongside quantitative measures. While prior work by Reviriego et al. (2024) highlights that ChatGPT-4 matches or exceeds human lexical diversity, they emphasize the need for further investigation into underlying patterns and explanations. The present paper addresses that call by examining a comprehensive corpus of 126 essays and extending the analysis to grammatical variation. The findings underscore AI's potential to produce text which resembles text written by humans while highlighting differences in variation and consistency.

References

Aubignat, Mickael, and Eva Diab (2023), Artificial intelligence and ChatGPT between worst enemy and best friend: The two faces of a revolution and its impact on science and medical schools, *Revue Neurologique* 179(6), 520–522. https://doi.org/10.1016/j.neurol.2023.03.004

Barrot, Jessie Saraza (2020), Integrating technology into ESL/EFL writing through Grammarly, *RELC Journal*, Article 3368822096663. https://doi.org/10.1177/0033688220966632

De Angelis, Luigi, Francesco Baglivo, Guglielmo Arzilli, Gaetano P. Privitera, Paolo Ferragina, Alberto E. Tozzi, and Caterina Rizzo (2023), ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health, *Frontiers in Public Health* 11, Article 1166120. https://doi.org/10.3389/fpubh.2023.1166120

Gul, Danish (2023), Rise of the machines: The potential threat of AI to engineering jobs, International Journal of Science and Research (IJSR) 12(7), 48–50. https://doi.org/10.21275/SR23620193839

Harari, Yuval Noah (2017), Reboot for the AI revolution, *Nature* (London) 550(7676), 324–327. https://doi.org/10.1038/550324a

Herbold, Steffen, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch (2023), A large-scale comparison of human-written versus ChatGPT-generated essays, *Scientific Reports* 13, Article 18617. https://doi.org/10.1038/s41598-023-45644-9

Leaver, Tama, and Suzanne Srdarov (2023), ChatGPT isn't magic: The hype and hypocrisy of generative artificial intelligence (AI) rhetoric, M/C Journal 26(5). https://doi.org/10.5204/mcj.3004

O'Neill, Ruth, and Alex Russell (2019), Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly, *Australasian Journal of Educational Technology* 35(1), 42–56.

Reviriego, Pedro, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández (2024), Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans, *Machine Learning with Applications* 18. https://doi.org/10.1016/j.mlwa.2024.10062

Rudnicka, Karolina (2023), Can Grammarly and ChatGPT accelerate language change? AI-powered technologies and their impact on the English language: wordiness vs. conciseness. *Procesamiento del Lenguaje Natural* 71, 205–214. https://doi.org/10.26342/2023-71-16

Skowron, Dominik, and Anna Bączkowska (2023), Assessment of various GPT models versus the human text: A quantitative analysis of lexis and cohesion, *Forum Filologiczne Ateneum* 1(11).

[1] Reviriego, P., Conde, J., Merino-Gómez, E. Gonzalo Martínez, J. H. (2024). Humans vs ChatGPT texts on TOEFL questions and HC3 dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.11199030.

Software and Data

Feinerer, Ingo, Kurt Hornik, and David Meyer (2008), tm: Text mining package (R package version 0.5-2). Available at: https://cran.r-project.org/package=tm

R Core Team (2023), R: A Language and Environment for Statistical Computing, Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/

Reviriego, Pedro, Javier Conde, Elena Merino-Gómez, and Gonzalo Martínez (2024), Humans vs ChatGPT texts on TOEFL questions and HC3 dataset [Data set], Zenodo. https://doi.org/10.5281/zenodo.11199030

When people overload their stomach(s): Non-verbal plural number agreement in Early and Late Modern medical discourse

Karolina Rudnicka and Richard Jason Whitt
(University of Gdansk, University of Nottingham)

This study investigates the use of singular and plural forms of *stomach* in Early and Late Modern English medical regimens, utilizing data from the Early Modern English Medical Texts (EMEMT) and Late Modern English Medical Texts (LMEMT) corpora. While modern English predominantly prefers plural number agreement between antecedents and coreferential terms (e.g. *Animals make their homes with the resources* vs. *Some animals make their home in it* (COCA)), as confirmed in the recent work by Rudnicka and Klégr (2023), the picture is more varied in other European languages (Rudnicka, 2024; Rudnicka and Klégr, 2024).

Given that the preference for distributive number in English has not yet been explored from a specific discourse domain or historical perspective, the present paper provides a timely addition to the literature. Specifically, we investigate whether the preference for the distributive plural was similarly applied in historical English medical texts. By addressing this question, the study contributes new insights to the relatively underexplored area of syntactic agreement in English (e.g., Halpert, 2016; Boeckx, 2006; Bondaruk et al., 2014; Keine, 2010).

The noun *stomach*[1] was chosen because it is a body part that everyone has only one of. Additionally, its compatibility with possessive pronouns like *their* and *ours* helps narrow the search to instances with plural subjects and identify third-person singular cases. Finally, it is the second-most frequently used noun in both corpora. To this end, we provide a statistical overview of the plural and singular forms of *stomach* across both corpora and explore the motivations behind the observed variations in medical discourse. Consider, for example, the following instances from the corpora:

1. (…) *we ought to take good hede we hurte nat our stomake by ouer* (EMEMT, 1528)
2. (…) *but to them who have weak stomacks* (EMEMT, 1656)
3. (…) *to those whose stomach and digestion is weak* (LMEMT, 1778)

Our findings reveal that, despite the modern preference for plural forms, at least one context – namely, the generalizing context – allows for *free variation*[2] between singular and plural forms of *stomach*, a pattern observable at least since the 16th century. Furthermore, our results show that the most common way of conveying generic meaning in this historical period and genre of regimens was through the definite article (*the*), while modern forms like singular *they* or other personal pronouns are either absent or rarely used in these texts.

References

Bondaruk, Anna, Grete Dalmi, and Alexander Grosu (eds) (2014), *Advances in the Syntax of DPs: Structure, Agreement and Case*, Amsterdam/Philadelphia: John Benjamins.

Boeckx, Cedric (2006), *Agreement Systems*, Amsterdam, Philadelphia: John Benjamins.

Brown, Keith, and Jim Miller (2013), *The Cambridge Dictionary of Linguistics*, Cambridge: Cambridge University Press.

Davies, Mark (2008–), *The Corpus of Contemporary American English (COCA): One Billion Words, 1990–2019*. Available at: https://www.english-corpora.org/coca/

Halpert, Claire (2016), *Argument Licensing and Agreement*, Oxford: Oxford University Press.

Keine, Stefan (2010), *Case and Agreement from Fringe to Core: A Minimalist Approach*, Berlin: Mouton de Gruyter.

Rudnicka, Karolina and Aleš Klégr (2023), Non-verbal plural number agreement. Between the distributive plural and singular: blocking factors and free variation, in K. Kopf and T. Weber (eds), *Free Variation in Grammar. Empirical and theoretical approaches*, Amsterdam/Philadelphia: John Benjamins, 74–98.

Rudnicka, Karolina (2024), Non-verbal plural number agreement – a pilot study comparing English and German using Oslo Multilingual Corpus data, *Slovo a slovesnost* 85(1), 27–54.

Rudnicka, Karolina and Aleš Klégr (2024), Non-verbal Plural Number Agreement in the Cross-linguistic Context: Combining Corpus Findings with Two Kinds of Acceptability Rating Results for English, German, Polish, and Czech, *Nordic Journal of English Studies* 23(2), 92–119.

Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr, and Jukka Tyrkkö (2010), *Early Modern English Medical Texts: The Corpus of Early Modern English Medical Writing*, Amsterdam/Philadelphia: John Benjamins.

Taavitsainen, Irma, Turo Hiltunen, Anu Lehto, Ville Marttila, Päivi Pahta, Maura Ratia, Carla Suhr, and Jukka Tyrkkö (2019), *Late Modern English Medical Texts: The Corpus*, Amsterdam/Philadelphia: John Benjamins.

The repertoire of English equivalents of the Lithuanian discourse particles *na/nu* and *va*: Evidence from the bidirectional English-Lithuanian parallel corpus

Anna Ruskan and Audronė Šolienė
(Vilnius University)

Discourse particles have been widely investigated in individual languages (Aijmer 2002; Haselow 2015; Heritage, Sorjonen 2018; Jasionytė-Mikučionienė 2019) as well as across languages (Degand 2014; Aijmer 2019) in terms of their multifunctionality, scope, clause periphery, categorial status, co-occurrences with other discourse elements as well as diachronic development (i.e. grammaticalization). The elusive multifunctional nature of discourse particles has been addressed in explorations of their translational correspondences in parallel corpora, which provide patterns and meanings of discourse elements which would be difficult to pin down by analyzing introspective data (Aijmer, Simon-Vandenbergen 2003; Usonienė et al. 2015). The present study aims to examine the diverse functional profile of the Lithuanian discourse particles *na/nu* and *va* by looking into their formal and functional English translational correspondences. The research method is a quantitative and qualitative contrastive analysis based on the data extracted from a self-compiled bidirectional parallel corpus – ParaCorp EN→LT→EN (Šolienė 2013).

The findings show that the formal and functional diversity of both particles is quite broad. However, it should be noted that the particle *na/nu* has more translational equivalents than the particle *va*. This variation in the number of translational equivalents could be attributed to the different semantic potential of the two particles and their advancement on the grammaticalization path. The particle *na*/*nu* has a broader general semantic meaning than *va* and thus receives numerous interpretations reflected in the number of translational correspondences. The most frequent equivalent of *na/nu* is the English particle *well*, which occurs in diverse dialogic contexts and performs various interpersonal and discourse structuring

functions. The particle *va* is frequently used as a demonstrative marker rendered in translation by the English deictic expressions *here* and *there*. The interpersonal functions of *va* are realised in translation by the English markers *see* and *look*, whereas the discourse structuring function is expressed by the particle *so*. The data obtained from the parallel corpus also allow to capture the functional potential of the particle combinations that the particles *na*/*nu* and *va* form, i.e. *na/nu va*, *tai va*, *nu tai va*, *o va*. The translational correspondences of the particle combinations reveal the types of integration (juxtaposition, addition, composition) (Cuenca, Crible 2019) that the particles display and their expressive force.

References

Aijmer, Karin (2002), *English Discourse Particles: Evidence from a Corpus,* Amsterdam/Philadelphia: John Benjamins.

Aijmer, Karin and Anne-Marie Simon-Vandenbergen (2003), The discourse particle *well* and its equivalents in Swedish and Dutch, *Linguistics* 41(6), 1123–1161.

Aijmer, Karin (2019), Challenges in the contrastive study of discourse markers: The case of *then,* in Ó. Loureda, I. R. Fernández, L. Nadal and A. Cruz (eds), *Empirical Studies of the Construction of Discourse*, Amsterdam/Philadelphia: John Benjamins Publishing Company, 17–42.

Jasionytė-Mikučionienė, Erika (2019), Subordinating conjunctions as discourse markers in Lithuanian, *Corpus Pragmatics*, 1–17.

Cuenca, María Josep and Ludivine Crible (2019), Co-occurrence of discourse markers in English: From juxtaposition to composition, *Journal of Pragmatics* 140, 171–184.

Degand, Liesbeth (2014), 'So very fast then': Discourse markers at left and right periphery in spoken French, in K. Beeching and U. Detges (eds), *Discourse Functions at the Left and Right Periphery: Crosslinguistic Investigations of Language Use and Language Change*, Brill: Leiden, 151–178.

Haselow, Alexander (2015), Left vs. right periphery in grammaticalization: The case of *anyway*. *New Directions in Grammaticalization Research*, edited by A. D. M. Smith, G. Trousdale and R. Waltereit, Amsterdam/Philadelphia: John Benjamins Publishing Company, 157–186.

Heritage, John and Marja-Leena Sorjonen (2018), *Between Turn and Sequence: Turn-Initial Particles Across Languages,* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Šolienė, Audronė (2013), *Episteminio modalumo ekvivalentiškumo parametrai anglų ir lietuvių kalbose*, (ms.), Humanitarinių mokslų daktaro disertacija, Vilnius: Vilnius University.

Usonienė, Aurelija, Audronė Šolienė, and Jolanta Šinkūnienė (2015), Revisiting the multifunctionality of the adverbials of ACT and FACT in a cross-linguistic perspective, *Nordic Journal of English Studies* 14(1), 201–231.

Categorising semantic prosody - from past discord to future consensus?

Mathias Russnes
(University of Oslo)

This paper investigates the stability of manual analysis of concordance lines and the inter-rater reliability of established methods of categorising semantic prosody. Semantic prosody is a concept associated with corpus linguistics, and describes the tendency of seemingly neutral items to occur in particular evaluative contexts (e.g. Sinclair 1996; Stewart 2010). In previous research (e.g. Louw 1993; Partington 2004), this tendency has often been categorised in binary terms, distinguishing between *positive* and *negative* prosodies. An example is the item *utterly,* which has been ascribed a negative prosody. However, this restricted system has also received criticism (e.g. Bednarek 2008), and certain researchers (e.g. Sinclair 1996; Stubbs 2001) have adopted a broader categorisation, more connected to a unit's semantic preference. For example, Sinclair (1996) argues that *true feelings* has a prosody of "reluctance" or "inability". Further, questions have also been raised regarding the impact of subjective judgements in the manual analysis performed to arrive at these categorisations, as well as the consistency of this approach

(e.g. Dilts and Newman 2006; Winter 2019). Previous inter-rater reliability studies in the related field of corpus-assisted discourse studies have yielded varying results (e.g. Marchi and Taylor 2009; Baker 2015), showing both convergence and divergence, indicating a lack of consistency across researchers in manual analysis. Yet, these issues have not been thoroughly explored within the context of semantic prosody, despite being briefly considered in certain studies, such as Zhang (2013). This paper aims to fill this research gap by addressing the stability of manual analysis and the reliability of the established methods of categorising semantic prosody through the following research question:

- Does a binary distinction between *positive* and *negative* or a more comprehensive categorisation using specific terms result in a higher inter-analyst consistency when determining semantic prosody?

To answer this research question, two experimental studies will be conducted, wherein four researchers analyse the same material independently using both methods. The material consists of a set of random concordance lines of the items *habit* and *views* gathered from the BNC2014, which have been chosen for their suspected tendencies of occurring in particular evaluative contexts. The studies will be performed in multiple rounds, and Krippendorff's Alpha will be used to assess the inter-rater reliability. The preliminary results indicate that a binary distinction between positive and negative offers a higher inter-analyst consistency than a more detailed categorisation. Further, although using specific terms can describe the wide range of pragmatic meanings expressed through semantic prosody more precisely, this approach may also obscure the borders between semantic preference and semantic prosody. By scrutinising the prevalent methods of the past, this paper aspires to enhance future studies of semantic prosody by solidifying the methodological framework used for categorisation. Additionally, the results could have wider implications for corpus linguistics more generally, where manual analysis remains a widely used approach (Winter 2019).

References

Baker, Paul (2015), Does Britain need any more foreign doctors? Inter-analyst consistency and corpus-assisted (critical) discourse analysis, in N., Groom, M., Charles and S., John (eds), *Corpora, Grammar and Discourse: In Honour of Susan Hunston*, Amsterdam: John Benjamins Publishing Company, 283–300.

Bednarek, Monika (2008), Semantic preference and semantic prosody re-examined, *Corpus Linguistics and Linguistic Theory* 4(2), 119–139.

Dilts, Philip and John Newman (2006), *Corpus Linguistics and Linguistic Theory* 2(2), 233–242.

Louw, Bill (1993), Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies, in M., Baker, G., Francis and E., Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, Amsterdam: John Benjamins Publishing Company, 157–176.

Marchi, Anna and Charlotte Taylor (2009), If on a winter's night two researchers… A challenge to assumptions of soundness of interpretation, *Critical Approaches to Discourse Analysis across Disciplines* 3(1), 1–20.

Partington, Alan (2004), 'Utterly content in each other's company': Semantic prosody and semantic preference, *International Journal of Corpus Linguistics* 9(1), 131–156.

Sinclair, John (1996), The search for units of meaning, reprinted in J., Sinclair and R., Carter (eds), *Trust the Text* (2004), London: Routledge, 24–48.

Stewart, Dominic (2010), *Semantic Prosody: A Critical Evaluation*, London: Routledge.

Stubbs, Michael (2001), *Words and Phrases: Corpus Studies of Lexical Semantics*, Oxford: Blackwell.

Winter, Bodo (2019), *Sensory Linguistics: Language, Perception and Metaphor*, Amsterdam: John Benjamins Publishing Company.

Zhang, Ruihua (2013), A corpus-based study of semantic prosody change: The case of the adverbial intensifier, *Concentric: Studies in Linguistics* 39(2), 61–82.

Using Large Language Models to enrich corpus metadata: The case of novels in the *Corpus of Historical American English*

Tanja Säily, Jukka Suomela, Florent Perek, Jimena Jiménez Real and Turo Vartiainen
(University of Helsinki, Aalto University, University of Birmingham)

Corpora are often conceptualized in terms of rich data vs. big data (e.g. Hiltunen et al. 2017). On the one hand, we have "small and tidy" corpora (Mair 2006) that include a rich array of metadata painstakingly compiled by hand, such as the 2.2-million-word *Parsed Corpus of Early English Correspondence* (2006). On the other hand, we have big and comparatively messier corpora like the 400-million-word *Corpus of Historical American English* (COHA; Davies 2010–) that allow the investigation of new kinds of research questions and less frequent phenomena, but this comes at the cost of granularity in their metadata, as compiling detailed metadata for these corpora would be too resource-intensive.

Modern machine learning shows great potential for enriching the metadata of big-data corpora (e.g. Öhman et al. 2019). We contribute to this trend by utilizing large language models (LLMs) to annotate the novels in the fiction section of COHA for subgenre and author metadata (8 subgenres, 3 age-based target audience categories, author birth year and gender). We hypothesize that this can be done by supplying the LLMs with a very limited amount of information (title, author, and publication year), as a great deal of metadata is present in sources like Wikipedia that are likely to be part of the training data for LLMs.

Our main goals are: (1) quantify how well LLMs can solve metadata annotation tasks; (2) develop best practices for accurate and resource-efficient metadata enrichment; (3) produce a metadata-enriched version of COHA; and (4) study if this metadata can shed light on phenomena observed in prior work.

We have manually annotated a random sample of 345 texts from COHA. This ground truth data set is then split into training, validation, and test data sets. We use the validation data set to compare the performance of different approaches: We quantify how much it helps if we include in the prompt some annotated training data points, or if we include in the prompt e.g. Wikipedia search results on the author name and the book title (a.k.a. retrieval-augmented generation). We compare different language models (especially gpt-4o and gpt-4o-mini), and different prompting strategies. We use the new "Structured Outputs" API by OpenAI to ensure that the output is always well-formed.

Once we have identified the most promising approach, we will use the test data set to estimate its accuracy on previously-unseen COHA texts and analyse the most common mistakes. Then we will use the same approach to annotate all of the COHA novels. We will make the LLM-produced metadata, as well as our ground truth data set and the relevant source code, freely available for download.

Finally, we will utilize this metadata to analyse gender differences in the productivity of the *way*-construction in COHA novels. Previous research (Perek et al. 2024) has found that the usage of men and women developed in tandem at first, but female authors began to use the construction significantly unproductively from the later twentieth century onwards. We will investigate whether this gender difference could be due to an imbalance in the subgenres of novels in COHA.

References

Davies, Mark (2010), *The Corpus of Historical American English: 400 Million Words, 1810–2009*. Available at: https://www.english-corpora.org/coha/

Hiltunen, Turo, Joe McVeigh, and Tanja Säily (2017), How to turn linguistic data into evidence?, in T. Hiltunen, J. McVeigh, and T. Säily (eds), *Big and Rich Data in English Corpus Linguistics: Methods and Explorations (Studies in Variation, Contacts and Change in English 19)*, Helsinki: VARIENG. Available at: https://urn.fi/URN:NBN:fi:varieng:series-19-0

Mair, Christian (2006), Tracking ongoing grammatical change and recent diversification in present-day Standard English: The complementary role of small and large corpora, in A. Renouf, and A. Kehoe (eds), *The Changing Face of Corpus Linguistics*, Amsterdam: Rodopi, 355–376.

Öhman, Emily, Tanja Säily, and Mikko Laitinen (2019), Towards the inevitable demise of everybody? A multifactorial analysis of -one/-body/-man variation in indefinite pronouns in historical American

English, *40th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 40)*, Neuchâtel, Switzerland, June 2019. Available at: https://tanjasaily.fi/talks/icame40_ohman_et_al_2019.pdf

Taylor, Ann, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen (2006), *Parsed Corpus of Early English Correspondence*, compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

Perek, Florent, Tanja Säily, and Jukka Suomela (2024), Historical sociolinguistics meets constructional change: Gender and the way-construction in the Corpus of Historical American English, *57th Annual Meeting of the Societas Linguistica Europaea (SLE 2024)*, Helsinki, Finland, August 2024. Available at: https://tanjasaily.fi/talks/sle57_perek_et_al_2024.pdf

Trump 2.0: The evolution of populist rhetoric from Twitter to Truth Social

Julia Schilling
(University of Bonn)

In an era dominated by digital communication, this study presents a nuanced analysis of Donald Trump's discourse in two distinct phases: his tweets from 2015 until his account was suspended following the Capitol attack in January 2021, referred to as "Trump 1.0," and his subsequent posts on Truth Social, referred to as "Trump 2.0," which began with the platform's launch on February 21, 2022. Through this comparative lens, I examine how Trump's strategic language choices embody populist rhetoric. Since announcing his candidacy in 2015, Trump has reshaped the political landscape with a confident style characterized by simplicity and directness, effectively appealing to disenfranchised voters disillusioned with conventional political discourse (Hawkins & Littvay, 2019).

While Trump is widely recognized as a populist leader, this study will explore the following research questions:

1. How does populist rhetoric manifest in Trump's communication strategies on Twitter and Truth Social?
2. In what ways do these rhetorical expressions evolve between the two platforms and across different electoral contexts, specifically contrasting the 2016 and 2020 elections with the 2024 election?

To answer these questions, this study uses a combination of qualitative and quantitative methods. The analytical framework is based on a multidimensional understanding of populism and focuses on the communicative opposition between "the pure people" and "the corrupt elite." Building on the framework created by Ernst et al. (2017), the analysis examines three key dimensions of Trump's discourse: people-centrism, anti-elitism, and popular sovereignty. First, I manually annotated a sample of 5,000 tweets and truth social posts by Trump from a corpus of approximately 31,000 tweets and truths, focusing on the key features related to populism identified in the study by Ernst et al. (2017), such as the feature *blaming the elite*. This annotated dataset was then used to train a machine learning model that predicts and classifies instances of populist rhetoric across the corpus to demonstrate how Trump's rhetoric constructs a narrative that not only empowers but also mobilizes his constituency.

Preliminary results indicate a strategic use of elements of populist rhetoric. Before each election, Trump's tweets and truths show consistent patterns in people-centered rhetoric, which then subside, an expected pattern in political communication. Following the 2020 election, there is a notable increase in anti-elitist rhetoric during the transition from "Trump 1.0" to "Trump 2.0," which continues with the platform shift. In "Trump 2.0," there is an increased prevalence of phrases such as "we will demolish the Deep State" and expressions like "America first," "under siege Second Amendment," and "gag order." These signal a reinforced call to resistance against perceived internal enemies and established structures, highlighting a narrative of threat and defense. This shift indicates a new strategic direction, targeting an audience specifically aligned with Truth Social, and underscores Trump's aim to emotionally engage and solidify his base.

References

Ernst, Nicole, Sven Engesser, and Frank Esser (2017), Bipolar populism? The use of anti-elitism and people-centrism by Swiss parties on social media, *Swiss Political Science Review* 23(3), 253–261.

Hawkins, Kirk, and Levente Littvay (2019), *Contemporary US Populism in Comparative Perspective*, in Frances E. Lee (ed.), *Elements in American Politics*, Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108644655

Project 2025 Unveiled: Keyness analysis and LIWC insights into political language

Julia Schilling and Robert Fuchs
(University of Bonn)

Situated within the field of political discourse analysis (Dunmire 2012), this study explores the linguistic strategies and rhetorical patterns employed in *Project 2025*, highlighting the role of think tanks in crafting politically influential documents that extend beyond electoral cycles. *Project 2025* is a comprehensive policy manifesto developed by the Heritage Foundation, aimed at influencing the structure and policy direction of an incoming Republican-led U.S. federal government led by Donald Trump. The new U.S. administration is widely expected to profoundly reshape American politics and, by dint of the U.S.'s economic and military stature, policies and politics around the world. Using corpus-based methods, including keyness and collocation analyses (Gabrielatos 2018, Brezina et al. 2015), alongside LIWC (Linguistic Inquiry and Word Count; Kahn et al. 2007) tools, we compare *Project 2025* with the Democratic and Republican Party platforms from 2016–2024 (amounting, in total, to just over 400,000 words). The analysis identifies distinct linguistic markers that position *Project 2025* as both a critique of current governance and a structured roadmap for future conservative policy implementation.

Keyness analysis reveals a focus on administrative restructuring, with terms like *agency*, *department*, and *office* occurring significantly more frequently in the *Project 2025* corpus than in either reference platform (see Fig. 1). These terms are often paired with action-oriented verbs such as *require*, *enforce*, and *prosecute*, emphasizing a focus on regulatory control and administrative efficiency. Adjectives such as *regulatory*, *executive*, and *conservative* highlight alignment with structural reform and ideological principles. The *Project 2025* manifesto also engages directly with contemporary political contexts, with collocations like *biden administration* and *trump administration* reflecting its positioning in response to recent U.S. governance.

The LIWC analysis offers additional insights into the psychological and rhetorical dimensions of the text. High Analytic Thinking scores indicate that *Project 2025* employs a systematic and logical style, consistent with its role as a detailed policy document. In contrast, its moderate Authenticity scores suggest a professional and detached tone, aimed at policymakers and thought leaders rather than the broader electorate. The manifesto's relatively low positive tone and heightened emphasis on security-related terms, such as *national security* and *intelligence*, underscore its defensive and prescriptive orientation.

Temporal focus analysis highlights a dual strategy: the manifesto references past conservative achievements, positioning itself within a historical ideological lineage, while also emphasizing a forward-looking vision through prescriptive policy language. For example, recurring phrases like *the next administration must* and *future reform efforts should* underscore its intent to influence governance structures proactively.

The findings of this study demonstrate how *Project 2025* leverages linguistic precision to present itself as a blueprint for conservative governance. Unlike the campaign-focused Democratic and Republican platforms, which prioritize voter engagement and immediate policy proposals, *Project 2025* is strategically crafted to articulate a cohesive and long-term vision for federal governance. Its detailed approach to administrative organization, emphasis on regulatory reform, and prescriptive tone position it as a critical document for understanding the evolution of conservative political discourse.

References

Brezina, Vaclav, Tony McEnery, and Stephen Wattam (2015), Collocations in context: A new perspective on collocation networks, *International journal of corpus linguistics* 20(2), 139–173.

Dunmire, Patricia L. (2012), Political discourse analysis: Exploring the language of politics and the politics of language, *Language and Linguistics Compass* 6(11), 735–751.

Gabrielatos, Costas (2018), Keyness Analysis: nature, metrics and techniques, in Taylor, C. C Marchi, A. (eds), *Corpus Approaches to Discourse: A critical review*. Oxford: Routledge, 225–258.

Kahn, Jeffrey H., Renée M. Tobin, Audra E. Massey, and Jennifer A. Anderson (2007), Measuring emotional expression with the Linguistic Inquiry and Word Count, *The American journal of psychology* 120(2), 263–286.

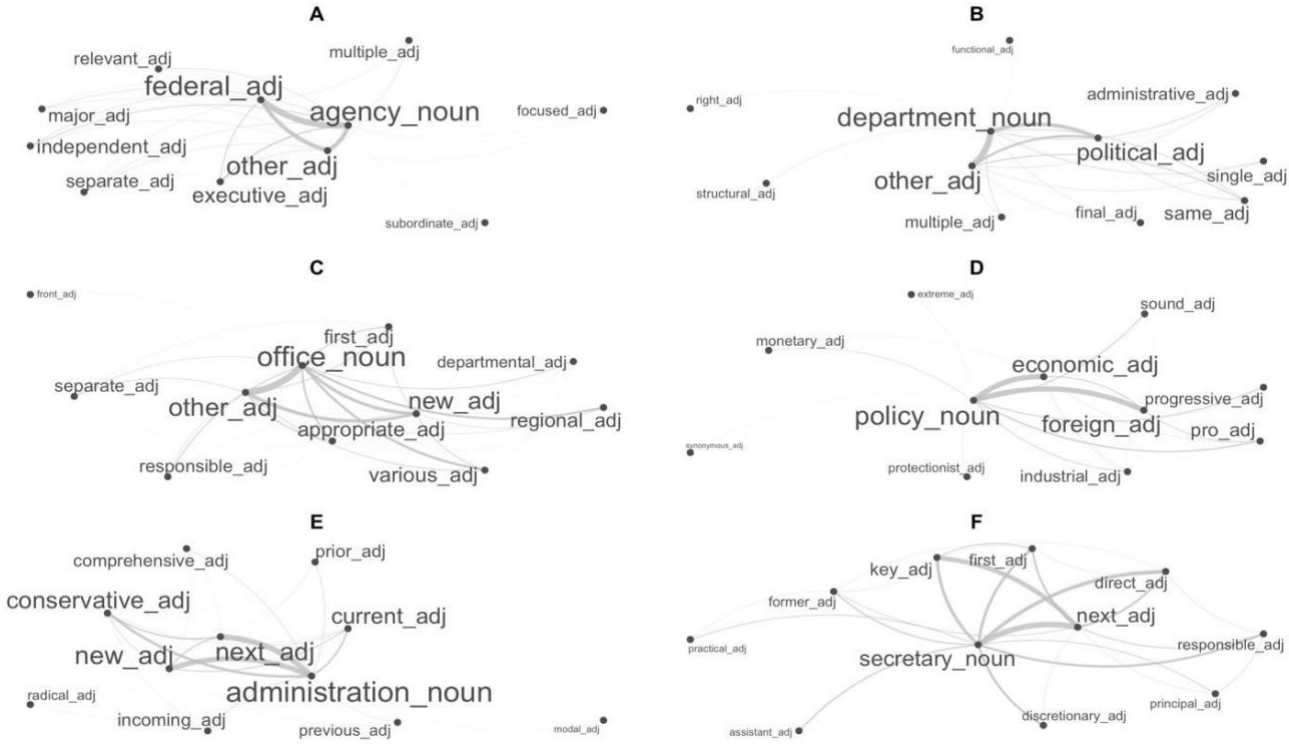


Figure 1. Top 10 Adjective Collocations of Top 6 Noun Keywords

The identification of a linguistic epicentre in Anglophone South Asia: Complementing corpus data with sociolinguistic evidence

Karola Schmidt
(Justus Liebig University Giessen)

The present study investigates the interplay between corpus and questionnaire data for research into linguistic epicentres. Generally, varieties of English are considered linguistic epicentres when they are endonormatively stabilised and display the potential to structurally influence neighbouring varieties (Hundt 2013, p. 185). In terms of linguistic evidence, Hundt argues that "statements about epicentric influence need to be based on language-use data […] as well as attitudinal data" (2013, p. 184). There is a substantial amount of research on linguistic epicentres overall (e.g. Bernaisch et al. 2022; Hoffmann et al. 2011; see Peters & Bernaisch 2022 for a recent overview) and Hundt et al. (2022) simulate the impact of attitudes on structural epicentral effects through agent-based modelling. Still, empirical insights from sociolinguistic

surveys have so far only played marginal roles in epicentral studies and have thus not been reconciled with corpus-based epicentral findings.

Using Anglophone South Asia, i.e. Bangladesh, India, the Maldives, Nepal, Pakistan and Sri Lanka, where Indian English has been profiled introspectively (Leitner 1992), but also empirically as a likely epicentre (Gries & Bernaisch 2016; Heller et al. 2017), I will present the type of complementary sociolinguistic evidence Hundt deems necessary for epicentre identification (2013, p. 184) by answering the following research questions:

1. From an attitudinal perspective, what is the likely linguistic epicentre of South Asian Englishes?
2. Can the structural evidence for epicentre identification in South Asian Englishes be reconciled with attitudinal evidence?

To this end, I show the results of a questionnaire study conducted in India and Sri Lanka in 2023 designed to explore the degree of endonormative stabilisation and epicentral status of Indian and Sri Lankan English as evidenced by respondents' attitudes towards their own and the respective other variety. The 566 thus compiled data points are treated with statistical modelling technique firmly rooted in corpus linguistics, that is a random forest analysis including sociobiographic meta data of the respondents as predictors.

The studies show matching results. The survey data also support the notion of Indian English as the strongest candidate for a South Asian epicentre. Adding such information on sociodemographic relationships between varieties shows the advantage of relying on complementary data sources to generate a full picture. Additionally, the similarity of methods employed in the structural explorations of a South Asian epicentre like, for example, Gries & Bernaisch (2016) and Heller et al. (2017) and then in this attitudinal study highlights the strength of corpus-linguistic methods for different types of data.

References

Bernaisch, Tobias, Stefan Th. Gries, and Benedikt Heller (2022), 'Theoretical Models and Statistical Modelling of Linguistic Epicentres', *World Englishes*, 41, 333–346. https://doi.org/10.1111/weng.12580

Gries, Stefan Th., and Tobias Bernaisch (2016), 'Exploring Epicentres Empirically: Focus on South Asian Englishes', *English World-Wide*, 37, 1–25. https://doi.org/10.1075/eww.37.1.01gri

Heller, Benedikt, Tobias Bernaisch, and Stefan Th. Gries (2017), 'Empirical Perspectives on Two Potential Epicentres: The Genitive Alternation in Asian Englishes', *ICAME Journal*, 41, 111–144. https://doi.org/10.1515/icame-2017-0005

Hoffmann, Sebastian, Marianne Hundt, and Joybrato Mukherjee (2011), 'Indian English – An Emerging Epicentre? A Pilot Study on Light Verbs in Web-Derived Corpora of South Asian Englishes', *Anglia*, 129, 258–280. https://doi.org/10.1515/angl.2011.083

Hundt, Marianne (2013), 'The Diversification of English: Old, New and Emerging Epicentres', in Daniel Schreier and Marianne Hundt (eds), *English as a Contact Language*, Cambridge: Cambridge University Press, 182–203. https://doi.org/10.1017/CBO9780511740060.011

Hundt, Marianne, Laetitia Van Driessche, and Dirk Pijpops (2022), 'Epicentral Influence via Agent-Based Modelling', *World Englishes*, 41, 377–399. https://doi.org/10.1111/weng.12584

Leitner, Gerhard (1992), 'English as a Pluricentric Language', in Michael Clyne (ed.), *Pluricentric Languages: Differing Norms in Different Nations*, Berlin/New York: Mouton de Gruyter, 179–237. https://doi.org/10.1515/9783110888140.179

Peters, Pam, and Tobias Bernaisch (2022), 'The Current State of Research into Linguistic Epicentres', *World Englishes*, 41, 320–332. https://doi.org/10.1111/weng.12581

Same, same, but, *erm*, *sort of* different?
Comparing fluencemes across Australian, British, Canadian, and New Zealand English

Karola Schmidt, Sandra Götz-Lehmann, Katja Jäschke and Stefan Th. Gries
(Justus Liebig University Giessen, University of Marburg, University of Siegen,
University of California, Santa Barbara)

Fluency in English has been widely researched in the past few decades, with previous studies having mainly taken psycholinguistic, cognitive or sociolinguistic perspectives (e.g., Albert 1980; Beier 2023; Crible 2018; Goldman-Eisler 1961; Tottie 2015). Studies indicate that speakers usually bridge the gap between online processing demands and speaking by using different kinds of "fluencemes" (Götz 2013: 8-9) such as filled pauses (such as *er*, *erm*), unfilled pauses (i.e. pauses that are not filled with a non-verbal sound), as well as discourse markers (such as *you know* and *like*, etc.), including "smallwords" (Hasselgren 2002) (such as *sort of*/*sorta*, *kind of*/*kinda*). The necessity to use such planning strategies has been widely recognised in earlier research on fluency (e.g. Beeching 2016; O'Connell & Kowal 2005). Despite the general consensus about the frequent existence of fluencemes in L1 English speech, contrastive studies investigating potential differences across different varieties of English have only been rarely conducted (see, however, Miller 2009 on discourse markers in Australian vs. New Zealand English, or Tottie 2011, 2015 on the use of filled pauses as "planners" in British vs. American English). Different studies on certain fluencemes based on sociolinguistic speaker variables, such as age or gender, have generated interesting results, however, so far, they mainly focused on one English variety (e.g. Fruehwald 2016; Laserna et al. 2014; Scheuringer et al. 2017; Sokołowski et al. 2020; Weiss et al. 2006; or the studies in Leuckert & Rüdiger 2021). While there is also a sizable amount of research on the use of individual fluencemes generally, we can observe a heavy focus on research into either discourse markers or filled pauses. To the best of our knowledge, contrastive research on combining different fluencemes across different varieties of English while also including sociolinguistic variation has not yet been conducted. Accordingly, we pose the following research questions:

1) Are there differences in fluenceme use between different L1 varieties of English?

2) Do sociolinguistic variables play a role in predicting the choice of particular fluencemes?

To answer these questions, this paper compares the use of three core fluencemes, i.e. discourse markers, filled pauses and unfilled pauses, across Australian, British, Canadian, and New Zealand English. These fluencemes were extracted and manually disambiguated from the private conversation sections of the respective components of the International Corpus of English (ICE-AUS, ICE-GB, ICE-CAN, and ICE-NZ). The data were normalized per speaker and linked with the sociobiographic metadata of the speakers. Analysis using random forests revealed a consistent fluenceme distribution across the four varieties, with unfilled pauses being the most common, followed by discourse markers, and then filled pauses. This pattern suggests a 'common fluenceme core' among L1 English varieties. The influence of sociolinguistic variables – gender, age, education, and occupation – was modest and exhibited diverse trends. Male speakers tend to use filled pauses more frequently but fewer unfilled pauses compared to female speakers. Increasing age did not significantly affect the frequency of these strategies. Both education and occupation showed a slight positive correlation with overall fluency.

References

Albert, Martin L. (1980), Language in normal and dementing elderly, in L. K. Obler and M. L. Albert (eds), *Language and Communication in the Elderly*, Lexington, MA: DC Heath and Co, 145–150.

Beeching, Kate (2016), *Pragmatic Markers in British English: Meaning in Social Interaction*, Cambridge: Cambridge University Press.

Beier, Eleonora J., Suphasiree Chantavarin, and Fernanda Ferreira (2023), Do disfluencies increase with age? Evidence from a sequential corpus study of disfluencies, *Psychology and Aging* 38(3), 203–218.

Crible, Ludivine (2018), *Discourse Markers and (Dis)fluency: Forms and Functions across Languages and Registers*, Amsterdam and Philadelphia: John Benjamins.

Fruehwald, Josef (2016), Filled pause choice as a sociolinguistic variable, in *Selected Papers from New Ways of Analyzing Variation 44* 22, 41–48.

Goldman-Eisler, Frieda (1961), A comparative study of two hesitation phenomena, *Language and Speech* 4(1), 18–26.

Götz, Sandra (2013), *Fluency in Native and Nonnative English Speech*, Amsterdam and Philadelphia: John Benjamins.

Hasselgren, Angela (2002), Learner corpora and language testing: Smallwords as markers of learner fluency, in S. Granger, J. Hung, and S. Petch-Tyson (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam and Philadelphia: John Benjamins, 143–173.

Laserna, Charlyn M., Yi-Tai Seih, and James W. Pennebaker (2014), Um ... who like says you know: Filler word use as a function of age, gender, and personality, *Journal of Language and Social Psychology* 33(3), 328–338.

Leuckert, Sven, and Sofia Rüdiger (2021), Discourse markers and world Englishes, *World Englishes* 40(4), 482–487.

Miller, Jim (2009), Like and other discourse markers, in P. Peters, P. Collins, and A. Smith (eds), *Comparative Studies in Australian and New Zealand English*, Amsterdam and Philadelphia: John Benjamins, 315–336.

O'Connell, Daniel C., and Sabine Kowal (2005), Uh and um revisited: Are they interjections for signaling delay?, *Journal of Psycholinguistic Research* 34(6), 555–576.

Scheuringer, Andrea, Ramona Wittig, and Belinda Pletzer (2017), Sex differences in verbal fluency: The role of strategies and instructions, *Cognitive Processing* 18(4), 407–417.

Sokołowski, Andrzej, Ernest Tyburski, Anna Sołtys, and Ewa Karabanowicz (2020), Sex differences in verbal fluency among young adults, *Advances in Cognitive Psychology* 16(2), 92–102.

Tottie, Gunnel (2011), Uh and um as sociolinguistic markers in British English, *International Journal of Corpus Linguistics* 16(2), 173–197.

Tottie, Gunnel (2015), Uh and um in British and American English: Are they words? Evidence from co-occurrence with pauses, in R. Torres Cacoullos, N. Dion, and A. Lapierre (eds), *Linguistic Variation: Confronting Fact and Theory*, New York and Abingdon: Routledge, 38–55.

Weiss, Elisabeth M., J. Daniel Ragland, Colleen M. Brensinger, Warren B. Bilker, Eberhard A. Deisenhammer, and Margarete Delazer (2006), Sex differences in clustering and switching in verbal fluency tasks, *Journal of the International Neuropsychological Society* 12(4), 502–509.

A tag is worth a thousand words? Evaluating AI image tagging for multimodal corpus construction

Hanna Schmück
(University of Augsburg, University of Glasgow)

Building on the work on multimodal corpora by McClure et al. (2011), Collins (2020), and especially Baker and Collins (2023), which served as a pilot study for this project, the present study evaluates the usability of an existing AI image tagging tool, Vertex AI (Google Cloud, 2023), for multimodal corpus construction. This avenue of research is exceptionally promising as it enables a broader exploration of the dynamics between words and images, as outlined by Nikolajeva and Scott (2000). Furthermore, it provides a framework for examining how images impact the perceived newsworthiness of events or facts, a concept discussed by Bednarek and Caple (2012) as well as Galtung and Ruge (1965).

This paper poses three critical research questions. Firstly, how can we meaningfully evaluate the performance of automatic image tagging tools such as Google Cloud's Vertex AI for corpus linguistic research? Secondly, in the context of a case study on the representation of Muslims in the British press, what level of accuracy does Vertex AI provide in tagging newspaper images? Thirdly, what tools will linguists need to effectively survey multimodal material, such as images, alongside or integrated in traditional corpus tools?

The corpus constructed for this project comprises articles from nine UK newspapers collected between December 2022 and December 2023 containing the terms "Muslim" or "Islam," with associated

images tagged via Vertex AI. The final corpus contains over 1.5 million tokens from 1,890 articles, 8,546 images, and 89,133 image tags. Images were pushed through the Vertex AI pipeline using a custom Python script, generating up to 50 tags per image alongside corresponding confidence scores. An initial pilot analysis of 100 randomly selected images identified several tagging inaccuracies. High-frequency tags were further scrutinized, with 540 images (15 images per high-frequency tag) evaluated by two independent raters (overall inter-rater agreement 0.79) to assess tag accuracy. Tags such as Building, Chin, Car, Human, and Protest showed high accuracy, while tags like Winter, Gesture, and FashionDesign were less reliable with an overall Precision of 0.54 for the 110 most frequent tags in the dataset.

A further finding indicates that some elements of the tagging process, such as the identification of faces, might be partially rule-based and especially error-prone. For instance, the Vertex AI confidence score for constituent elements of a face (e.g., Nose, Chin) was identical across the entire dataset and did not account for cases where parts of a face were obscured. The iterative refinement process, which involved removing unreliable or irrelevant tags, increased overall tag accuracy to approximately 90%. To address research question three, the paper also presents the Image Tag Explorer tool which was developed to view images by tags or adjacent words in the corpus in the hopes of facilitating effective research of multimodal corpora.

This paper underscores the transformative potential of multimodal corpus linguistics in understanding complex discourse patterns. The inclusion of image tags provided profound insights into the representation of Islam and Muslims in the context of our case study, enriching traditional text-based analysis. Future directions include integrating custom tags and improving concordance presentation, ensuring the robustness and credibility of multimodal corpus approaches.

References

Bednarek, Monika, and Helen Caple (2017), *The Discourse of News Values: How News Organizations Create Newsworthiness*, Oxford: Oxford University Press.

Baker, Paul, and Luke Collins (2023), Creating and analysing a multimodal corpus of news texts with Google Cloud Vision's automatic image tagger, *Applied Corpus Linguistics* 3(1), 100043. https://doi.org/10.1016/j.acorp.2023.100043

Collins, Luke C. (2020), Working with images and emoji in the Dukki Facebook Corpus. In: Rüdiger, S., and D. Dayter (eds), *Corpus Approaches to Social Media*, Amsterdam: John Benjamins, 175–196.

Galtung, Johan, and Mari Holmboe Rüge (1965), The structure of foreign news, *Journal of Peace Research* 2(1), 64–91.

Google Cloud (2023), Vertex AI (Version 1.0) [Software], Google.

McClure, Kimberly J., Rebecca M. Puhl, and Chelsea A. Heuer (2011), Obesity in the news: Do photographic images of obese persons influence antifat attitudes?, *Journal of Health Communication* 16(4), 359–371.

Nikolajeva, Maria, and Carole Scott (2000), The dynamics of picturebook communication, *Children's Literature in Education* 31(4), 225–239.

Exploring the past via archival corpora: The People's Collection of Wales Corpus

Hanna Schmück and Marc Alexander
(University of Augsburg, University of Glasgow)

Work-In-Progress

This work-in-progress paper investigates the potential of utilising archival materials as corpora. Despite the abundance of archival collections across the UK, including institutional and community archives, these resources remain underappreciated and underutilised in corpus linguistic research and the cultural value of language as a heritage object is not sufficiently recognised. The aforementioned collections do, however,

offer access to unique and diverse linguistic varieties that are difficult to source elsewhere and often encompass a wide range of subjective and politically charged viewpoints (Fitzgerald, 2022). Inspired by Pagenstecher and Pfänder (2017), this paper emphasises the benefits of stronger cooperation between historians and linguists and proposes a practical approach to strengthening this relationship via creating archival corpora.

The approach in this short paper stands out from existing work on transforming historical oral testimonies into spoken corpora (see Clary-Lemon, 2010; Fitzgerald, 2022) in that it focuses on constructing corpora based on secondary textual information such as metadata from large archival collections. We pose two core research questions: What can corpora reveal about communities of practice? How can we transform archival data into corpora? We also offer some initial thoughts on keyness analyses as tools to extract cultural differences in corpora of this type.

To explore these questions, we conduct a case study transforming the archival metadata from the People's Collection Wales (PCW) into a corpus containing over 7.7 million tokens from over 153,000 individual metadata files. This metadata includes a wide range of descriptions of community-generated digital content, such as scans of photographs, transcriptions of letters, interviews, and videos. Our paper details the practical steps of identifying, collecting, and cleaning archival data for corpus research, and illustrates the unique insights that can be uncovered through the use of such archival materials, compared to existing corpora and historical narratives.

The study aims to create a resource that provides profound insights into local communities of practice and cultural differences, and highlights language as a formative part of culture. Examining archival corpora like the PCW corpus has immediate applications in sociolinguistics, dialectology, and narrative discourse analysis (Roller, 2015), with further avenues for exploration. Possible research questions relating to this dataset would be e.g. whether – and if so how – the language in community archives systematically differs from language in institutional archives with regards to emotivity. In looking at the past through the lens of archival metadata, this research outlines a future pathway for corpus linguists that aims for a more community-based, culturally aware, and interdisciplinary approach to language.

References

Clary-Lemon, Jennifer (2010), '"We're Not Ethnic, We're Irish!": Oral Histories and the Discursive Construction of Immigrant Identity', *Discourse & Society*, 21(1), 5–25. https://doi.org/10.1177/0957926509345066

Fitzgerald, Christopher (2022), *Investigating a Corpus of Historical Oral Testimonies: The Linguistic Construction of Certainty*, 1st ed., London: Routledge. https://doi.org/10.4324/b22799

Pagenstecher, Cord, and Stefan Pfänder (2017), 'Hidden Dialogues: Towards an Interactional Understanding of Oral History Interviews', in Erich Kasten, Katja Roller, and Joshua Wilbur (eds), *Oral History Meets Linguistics*, 185–207.

Roller, Katja (2015), 'Towards the "Oral" in Oral History: Using Historical Narratives in Linguistics', *Oral History*, 43(1), 73–84.

Of machine learning and witch trial papers: Insights into Historical English Corpora

Christa Schneider
(University of Bern)

"Per Corpora ad Astra: Exploring the Past, Mapping the Future," perfectly encapsulates the goals of my research, which bridges computational tools and historical linguistics to investigate early English corpora, as e.g. outlined in Ehrmann et al. 2023 or Schweter and Bitzl 2019.

By applying machine learning techniques to Scottish Witch Papers and Witch Trial Papers from Salem (USA), this study explores how linguistic and cultural dimensions of early English can be analysed using innovative, data-driven approaches. Premodern English presents a variety of challenges to linguistic

analysis which are particularly acute when applying machine learning models, as these are often designed for modern languages and fail to account for the complexities of historical texts (see Schneider, Lehmann and Schneider 2015). My study extends previous work by explicitly adapting Information Extraction (IE) techniques – specifically Named Entity Recognition (NER) and Sentiment Analysis (SA) – to overcome these challenges.

Research Questions:

My research addresses two key questions:

1. Is it possible to adapt machine learning tools, particularly Information Extraction (IE) methods such as Named Entity Recognition (NER) and Sentiment Analysis (SA), to account for the specific challenges of historical English texts (see Ehrmann et al. 2023)?
2. What insights into the linguistic and socio-cultural contexts of early English can these methods provide, especially regarding naming conventions and the emotional tone of the witch trial corpora (see Hundt, Denison and Schneider 2021; Koncar, year unknown)?

Approach and Data:

My study focuses on Scottish Witch Papers from the 16th and 17th century, and on Witch Trial Papers from Salem (USA) from the late 17th century. These texts are rich in linguistic and cultural data but present challenges such as orthographic variation, sparse data, and semantic shifts over time that necessitate a custom-tailored computational approach. By integrating adaptations to traditional NER and SA methods this work offers a concrete contribution to our understanding of early English corpora.

Methods:

NER is applied to extract personal names, enabling a diachronic analysis of naming strategies (see Ehrmann et al. 2023; Schweter/Bitzl 2019). In the Scottish Witch Papers, this method demonstrates how naming conventions reflect and reinforce societal hierarchies and norms in early modern Scotland. The Salem documents, by contrast, provide a perspective on linguistic variations shaping social identity (see also Schneider 2022). Sentiment Analysis quantifies and visualizes trends in subjective language, particularly sentiments associated with fear and accusation during moments of crisis. By recalibrating modern sentiment lexicons to suit the historical context, the present study interrogates the efficacy of contemporary SA tools when applied to premodern data. Furthermore, it provides specific evidence of emotional peaks correlated with critical events in witch trials, adding depth to our understanding of cultural dynamics during these periods.

Expected Results:

NER reveals nuanced patterns in the use of names, demonstrating how language reflected and reinforced societal structures. SA highlights emotional trends and subjective language during pivotal moments in witch trials, offering insights into the cultural and emotional landscapes of the period. These concrete outcomes demonstrate that the proposed methods not only overcome limitations of standard tools but also offer new perspectives on linguistic shifts and social dynamics in early English societies.

Conclusion:

This presentation underscores the value of integrating computational tools into corpus-based historical sociolinguistics. By adapting machine learning techniques to the complexities of historical English, this study showcases how exploring historical corpora can map new directions for linguistic and cultural research.

References

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet (2023), Named Entity Recognition and Classification in Historical Documents: A Survey, *ACM Computing Surveys* 56(2), 1-47. Available at: https://dl.acm.org/doi/10.1145/3604931 [accessed 16 April 2025].

Hundt, Marianne, David Denison, and Gerold Schneider (2012), Retrieving Relatives in Historical Corpora, *Literary and Linguistic Computing* 27(1), 3–16. Available at: https://www.zora.uzh.ch/id/eprint/52961/5/Retrieving_relatives_final_LLC.pdf [accessed 16 April 2025].

Koncar, Philipp, Bernhard C. Geiger, Christina Glatz, Elisabeth Hobisch, Sanja Sarić, Martina Scholger, Yvonne Völkl, and Denis Helic (2022), A Sentiment Analysis Tool Chain for 18th Century Periodicals, in *Fabrikation von Erkenntnis: Experimente in den Digital Humanities.* Available via Melusinapress.lu, https://www.melusinapress.lu/read/ezpg-wk34/section/01ef33d1-8d9d-4391-a488-7d090f719858 [accessed 16 April 2025].

Schneider, Gerold (2022), Systematically Detecting Patterns of Social, Historical and Linguistic Change: The Framing of Poverty in Times of Poverty, *Trans Philologic Soc* 120, 447–473. Available on Wiley Online Library at: https://onlinelibrary.wiley.com/doi/10.1111/1467-968X.12252 [accessed 16 April 2025].

Schneider, Gerold, Lehmann Hans Martin, and Peter Schneider (2015), Parsing early and late modern English corpora, *Digital Scholarship in the Humanities* 30(3), 423–349, Oxford University Press: doi:10.1093/llc/fqu001

Schweter, Stefan and Marcus Bitzl (2019), Towards Robust Named Entity Recognition for Historic German, *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 96–103. Available on GitHub: https://github.com/dbmdz/historic-ner [accessed 16 April 2025].

And scald him (but not too much)
Negative directive speech acts in Middle English culinary recipes

Ulrike Schneider
(Johannes-Gutenberg-Universität Mainz)

Culinary recipes are full of directives, i.e. speech acts which propose a future course of action. The grammatical constructions used in such speech acts in modern recipes tend to fall into one of three categories, a) result-oriented/impersonal, such as the infinitive in (1), b) action-oriented, such as the imperative in (2), or c) preparation- oriented/cooperative, such as in (3). Which options are permissible depends on the language as well as on stylistic factors (cf. Brdar-Szabó C Brdar 2009).

1. *Teig* ***einfüllen*** *und sofort* ***backen***. (*Das elektrische Kochen*, 2000) dough in-fill-INF and immediately bake-INF (German)
2. ***Put*** *in the oven for 10 minutes*. (*How to Be a Domestic Goddess*, 2003)
3. ***Retiramos*** *del horno y* ***dejamos*** *enfriar*. (centrallecheraasturiana.es) Remove-3PL of.the oven and let-3PL cool (Spanish)

Taavitsainen (2001) and Grund (2003) note that the medieval predecessors of the modern recipe also alternated between a "more involved" or "reader-oriented" and a "more detached" or "substance-centred" way of giving advice. It is therefore surprising that the general tenor is that the syntax of historical cooking recipes – or *receipts* as they would have been called – is "simple" and "repetitive" (Carroll 1999; Arendholz et al. 2013; Diemer 2013), displaying "rarely any degree of complexity beyond temporal clauses" (Görlach 1992: 749) and that verbs almost exclusively occur in the imperative (cf. e.g. Görlach 1992; Bator C Sylwanowicz 2018).

The present paper presents a quantitative investigation of Middle English culinary recipes which explores the degree of grammatical variation in this text type. It focusses on negative directive speech acts, i.e. on contexts in which cooks are instructed on what not to do, as in (4).

4. *And raysyns of corauns forȝete noȝt* (MS Sloane 1986, ME culinary) and raisins of currants forget not

In order to not only be able to tell *which* grammatical options were used by authors of Middle English culinary recipes, but also to determine whether they were characteristic of this genre and period, the study uses three self-compiled corpora comprised of Middle English (ME) medicinal recipes, Middle English (ME) culinary recipes and Early Modern English (EModE) culinary recipes with a total size of over 190,000 words. These contain 426 tokens of negative directives which were then categorised in terms of grammar and discourse function.

The results show that while the imperative predominates, authors make use of a variety of other grammatical options, yet none that fall into the class of cooperatives, i.e. none with first-person subjects. On the impersonal end of the scale, the culinary recipes even contain fixed elliptical expressions without a verb, such as the one in (5).

5. *And no sauce but salt*. (MS Harley 4016, ME culinary)

Overall, ME culinary recipes are characterised by their large number of negative directives and specifically by directives functioning as degree modifiers (as in the title of this paper), which appears to be a compensation strategy for the dearth of numeric quantification in culinary recipes of this period.

References

Arendholz, Jenny, Wolfram Bublitz, Monika Kirner, and Iris Zimmermann (2013), 'Food for Thought – or, What's (in) a Recipe? A Diachronic Analysis of Cooking Instructions', in Cornelia Gerhardt, Maximiliane Frobenius, and Susanne Ley (eds), *Culinary Linguistics: The Chef's Special*, Amsterdam/Philadelphia: John Benjamins, 119–137.

Bator, Magdalena, and Marta Sylwanowicz (2018), 'Early English Recipes – Development of the Text Type', *Academic Journal of Modern Philology*, 7, 29–54.

Brdar-Szabó, Rita, and Mario Brdar (2009), 'Indirect Directives in Recipes: A Cross-Linguistic Perspective', *Łódź Papers in Pragmatics*, 5(1), 107–131.

Carroll, Ruth (1999), 'The Middle English Recipe as a Text-Type', *Neuphilologische Mitteilungen*, 100(1), 27–42.

Diemer, Stefan (2013), 'Recipes and Food Discourse in English – A Historical Menu', in Cornelia Gerhardt, Maximiliane Frobenius, and Susanne Ley (eds), *Culinary Linguistics: The Chef's Special*, Amsterdam/Philadelphia: John Benjamins.

Görlach, Manfred (1992), 'Text Types and Language History: The Cookery Recipe', in Matti Rissanen, Ossi Ihalainen, Terttu Nevalainen, and Irma Taavitsainen (eds), *History of Englishes: New Methods and Interpretations in Historical Linguistics*, Berlin: Mouton de Gruyter, 736–761.

Grund, Peter (2003), 'The Golden Formulas: Genre Conventions of Alchemical Recipes in the Middle English Period', *Neuphilologische Mitteilungen*, 104(4), 455–475.

Taavitsainen, Irma (2001), 'Middle English Recipes: Genre Characteristics, Text Type Features and Underlying Traditions of Writing', *Journal of Historical Pragmatics*, 2(1), 85–113.

The development of modal verbs of strong obligation in Late Modern British and American English (1750–2000)

Ole Schützler
(Leipzig University)

Twentieth-century developments in the modal verb systems of British and American English (BrE and AmE) have collectively been dubbed 'modal decline' (e.g. Leech 2013; Smith 2003; Leech et al. 2009), which describes a decrease in the frequencies of core modals partly compensated by higher frequencies of semi-modals. Within the semantic domain of strong obligation, for example, the frequency of MUST (ex. 1) has been decreasing, while frequencies of the verbal predicates HAVE TO (ex. 2) and NEED TO (ex. 3) have been increasing (Millar 2009; Leech 2013). Other relevant constructions involve (HAVE) GOT TO (including forms with *gotta*), as in example (4). Two broad, compatible processes may be responsible for these developments: grammaticalization (Hopper & Traugott 2003) and democratization/colloquial- ization (Farrelly & Seoane 2012).

1. Courtesying humourously, I **must** beg your pardon for that, my dear. (ARCHER, 1796sarg_d4a)

2. [This] ought to remind us how much we **have to** answer for, if we neglect our duty. (ARCHER, 1836marr_f5b)

3. She **needs to** have a routine. (ARCHER, 1964gelb_d8a)
4. The rent **has got to** be paid. (ARCHER, 1897cran_f6a)

This paper traces changing preferences regarding the marking of strong obligation back into the Late Modern English period, inspecting the relative frequencies of relevant verbs in both American and British English. The expectation is that, even prior to the 20th century, AmE is at the vanguard of the change.

Using written data from the ARCHER corpus (v. 3.2; Yánez-Bouza 2011), the frequencies of the verbs MUST, HAVE TO, NEED TO and (HAVE) GOT TO are compared for the years 1700– 1998. The analysis includes eight genres that are available in both varieties for the entire period: diaries, drama, fiction, journals, legal writing, letters, news, and science. A total number of *n* = 3,018 occurrences (in *n* = 791 texts) is analysed with multinomial mixed- effects regression models implemented with the R-package brms (Bürkner 2020). Separate models with the four-level outcome variable 'verb' are run for AmE and BrE. Fixed-part predictors are year of publication and the author's sex; random intercepts are specified for genre and text.

Results suggest that, right from the start of the period, the frequencies of all four verbs are changing in the expected direction, with MUST losing ground and the other three verbs becoming more frequent. American English is generally more dynamic, either reflecting the change earlier, or at a higher rate of change. Apart from discussing the plausibility of these findings in terms of grammaticalization and democratization, it is argued that they constitute a case of *colonial innovation* (Hundt 2009) in the widest sense.

References

Bürkner, Paul-Christian (2020), *brms: Bayesian Regression Models Using Stan*, R package version 2.16.1. Available at: https://cran.r-project.org/web/packages/brms/brms.pdf

Farrelly, Michael, and Elena Seoane (2012), 'Democratization', in Terttu Nevalainen and Elizabeth Closs Traugott (eds), *The Oxford Handbook of the History of English*, Oxford: Oxford University Press, 392–401.

Hopper, Paul J., and Elizabeth Closs Traugott (2003), *Grammaticalization*, Cambridge: Cambridge University Press.

Hundt, Marianne (2009), '"Colonial Lag", "Colonial Innovation", or Simply "Language Change"?', in Günter Rohdenburg and Julia Schlüter (eds), *One Language, Two Grammars? Differences Between British and American English*, Cambridge: Cambridge University Press, 13–37.

Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith (2009), *Change in Contemporary English: A Grammatical Study*, Cambridge: Cambridge University Press.

Leech, Geoffrey (2013), 'Where Have All the Modals Gone? An Essay on the Declining Frequency of Core Modal Auxiliaries in Recent Standard English', in Juana I. Marín-Arrese, Marta Carretero, Jorge Arús Hita, and Johan van der Auwera (eds), *English Modality: Core, Periphery and Evidentiality*, Berlin: Mouton de Gruyter, 95–115.

Millar, Neil (2009), 'Modal Verbs in TIME: Frequency Changes 1923–2006', *International Journal of Corpus Linguistics*, 14(2), 191–220.

Smith, Nicholas (2003), 'Changes in the Modals and Semi-Modals of Strong Obligation and Epistemic Necessity in Recent British English', in Roberta Facchinetti, Manfred Krug, and Frank R. Palmer (eds), *Modality in Contemporary English*, Berlin: Mouton de Gruyter, 241–266.

Yáñez-Bouza, Nuria (2011), 'ARCHER Past and Present (1990–2010)', *ICAME Journal*, 35, 205–236.

A corpus-based analysis of general extenders in Irish English

Martin Schweinberger
(The University of Queensland)

This study examines the use of general extenders (cf. 1 and 2) from a corpus-based variationist perspective based on the Irish component of the *International Corpus of English* (ICE-Ire) (Kallen & Kirk 2007). The

analysis aims at determining if the general extender system of IrE is undergoing change as has been described in other varieties (see, e.g., Tagliamonte & Denis 2010 for Toronto English). To this end, the analysis uses multivariate statistics to investigate the existence of linguistic constraints and the potential social stratification of general extender use.

1. Adjunctive general extenders

a. She 's her bag packed and <,> washed out and <,> her pencil case filled **and so on** (ICE-Ire: S1A-001$B)

b. But because it 's a real Protestant thing they were never treated with the right respect **and all this sort of stuff** (ICE-Ire: S1A-005$A)

c. Sure people 're bringing in fruit for people **and stuff** (ICE-Ire: S1A-005$A)

2. Disjunctive general extenders

a. You 'd think it was hot **or something** <&> laughter </&> (ICE-Ire: S1A-003$B)

b. I don't know what he was writing whether it was the cars **or whatever** you know (ICE-Ire: S1A-012$A)

From a sociolinguistic perspective, general extenders are particularly interesting as they

1. are semantically bleached and are thus bias towards attracting social meanings
2. occur frequently in informal conversational data
3. can be easily extracted and
4. occur in well-circumscribed variable contexts.

Structurally, general extenders can be divided into adjunctive (*and* X) and disjunctive (*or* X) general extenders (see Overstreet 2020) and they have been described as consisting of four main components (Tagliamonte & Denis 2010):

i. Connector: e.g., *and*, *or*
ii. Quantifier: e.g., *all*, *every*, *some*
iii. Generic: e.g., *stuff*, *thing(s)*, *shit*
iv. Comparative: e.g., *like that*, *of that type*

The results of this study show that *and all* (182), *or something* (157), *or whatever* (95), *and everything* (77), *and so on* (68), and *and stuff* (67) are the most frequent general extenders in Irish English, jointly comprising 66.5 percent of all instances. The results of a conditional inference tree (or CART analysis) show that the use of general extenders is indeed socially stratified and differs across registers:

- *or something* is both strongly preferred by older speakers in more formal contexts such as scripted speeches and by younger speakers in informal conversation
- *and all* is preferred by younger speakers in public settings

The results are interpreted to show moderate change across social dimensions but, more importantly, they show that general extenders are perceived as being genre specific and thus rather serve as markers of formality than as social identity indices.

References

Kallen, Jeffrey C. and John Kirk (2007), ICE-Ireland: Local variations on global standards, in Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl (eds), *Creating and digitizing language corpora*, 121–162, London: Palgrave Macmillan.

Overstreet, Maryann (2020), The English general extender: The forms and functions of a new linguistic category, or something, and stuff, *English Today* 36(4), 47–52.

Tagliamonte, Sali and Derek Denis (2010), The Stuff of Change: General Extenders in Toronto, Canada, *Journal of English Linguistics* 38(4), 335–368.

The English loanwords: An addition or a subtraction to an Indian Language

Gatha Sharma
(Shiv Nadar Institution of Eminence)

English is a language of prestige in India. The imposition of English as the 'official language' of India, from 1835 to 1947, by British colonial regime made English language the 'Lingua Franca' of India. Its wide-spread usage in the legislative, executive and judicial branches of Indian government, pre- and post-Independence of India, has made it the most important language among all Indian languages. The knowledge of this language is a pre-requisite in many professional fields in India today. Popham (1996) sums up this phenomenon well when he says, "While the engine of colonialism long ago ran out of steam, the momentum of its languages is still formidable" (qtd. in Master,1998, p. 717).

India has amazing linguistic diversity. The People's Linguistic Survey of India (PLSI), conducted from 2009 to 2013, has identified, and recorded 780 Indian languages. All Indian languages have English loanwords. Borrowing is a process by which one language or dialect takes or incorporates some linguistic elements from another (Arlotto, 1972). Many English loanwords have also adapted themselves according to the grammatical markings and sounds of the native Indian languages. This corpus-based study aims to capture the most common English loanwords used by the speakers of Hindi, an Indian language, and to examine whether the code-mixing of English loan words in Hindi language is an addition to the richness of Hindi language or leads to lexical depletion in the language.

Methodology- Hindi grammar has a specific category for the 'loanwords', named as videshaj. Hindi dictionaries have comprehensive lists of loanwords, borrowed from different languages. The first corpus for the study is created from three Hindi language dictionaries- Hindi Shabdasagara, Vardha Hindi Shabdakosh, and Nalanda Vishal Hindi Shabdasagara. These dictionaries are the most prominent dictionaries of Hindi language and are available in digitized form as well. The lists of English loanwords are extracted from the dictionaries manually. The second corpus for this study is comprised of 100 copies of three most read Hindi Newspapers -Dainik BhaskR, Dainik Jagaran and Amar Ujala, collected from Jan. 1, 2025 to Feb. 3, 2025. Newspapers write in a language which is used by common people in their daily life. Thus, the newspapers are the best medium to find out the most common lexicon (including English loanwords) used by native speakers. The digital versions of newspapers are used to extract English loan words through ctrl+F function. The first corpus is used to determine-a) the total number of English loanwords used by Hindi speakers; and b) the existence of same or similar meaning words in Hindi language. The second part of the study is to compare the usage of English loanwords and their counterparts in Hindi language in Hindi newspapers. The comparative study will determine the total number of English loanwords in common usage; their word categories; and the context in which they are used. The study will also examine whether adaptation of English loanwords leads to lexical depletion in the native language or not.

References

Arlotto, Anthony (1972), *Introduction to Historical Linguistics*, New York: University Press of America.

Government of India (n.d.), *Language Census Reports*, available at: https://language.census.gov.in/map/data/showLSIReports

Master, Peter (1998), 'Positive and Negative Aspects of the Dominance of English', *TESOL Quarterly* 32(4), 716–727.

Differentiating translated and non-translated English diplomatic discourse: A diachronic corpus-based elastic net analysis of the United Nations General Debate (1946–2022)

Lin Shen and Haidee Kotze
(University of Cambridge, Utrecht University)

Extensive progress has been made in computational studies aiming to differentiate translated and non-translated language (with a particular emphasis on English), and to identify and explain the causes of the distinctive features of translated language (e.g., Hu & Kübler, 2021; Rabinovich & Wintner, 2015). Nevertheless, several empirical research gaps remain. This study addresses two of these. First, the application of statistical methods to address multicollinearity remains underexplored in this area of research. Elastic net regularization (Zou & Hastie, 2005) offers a robust solution for managing multicollinearity and improving the interpretability of the resulting models (Herawati et al., 2024). Second, there is a notable lack of diachronic studies that trace changes in translation features over extended periods. Addressing this gap could help understand the interplay between translation and language change, as translation may initiate, reflect, or constrain the change in the target language (House, 2008).

This paper seeks to address these research gaps by examining the linguistic features that differentiate translated from non-translated English diplomatic discourse at the United Nations General Debate (UNGD) over the period spanning 1946 to 2022. Using a corpus-driven approach, the study aims to answer two research questions: 1) Which linguistic features (if any) set apart translated and non-translated English diplomatic discourse at the UNGD? How do they relate to existing research on the features of translated language? 2) If there is a set of such distinctive translation features in English UNGD discourse, do these features change over time? How do these changes in translated discourse (if any) coincide with, precede, or lag behind the general processes of (contact-influenced) language change in English?

The corpus used in this study consists of a 1,393,720-token corpus of translated English UNGD discourse and a 1,381,819-token corpus of non-translated English UNGD discourse. The details of the corpus for time periods and source languages (of the translated texts) are presented in the supplementary materials (https://osf.io/8p9vn/?view_only=9e4e02387a9c40f493ff17e678962965). The analysis incorporates 245 linguistic features (listed in the supplementary materials), derived from Multidimensional Analysis (Biber, 1988; Nini, 2019), the L2 Syntactic Complexity Analyzer (Lu, 2010, 2017), and the Linguistic Feature Toolkit (Lee & Lee, 2023) to comprehensively assess textual properties. To answer RQ 1, elastic net analysis is used to select the most predictive indicators of translated English from the parameters. To examine change over time (RQ 2), mixed effects modeling is used with the lme4 package (Bates et al., 2015). The diachronic changes are visualized in scatter plots with ggplot2 (Wickham, 2016).

The translation features selected by the elastic net regression suggest that translated English UNGD texts exhibit higher levels of formality, abstractness, and informational density than non-translated English UNGD texts, aligning with existing studies on translated English (Kruger & van Rooy, 2016, 2018; Xiao, 2009). Over time, the translated texts tend to lag behind non-translated speeches in adopting language changes in the translation features, supporting the notion that translation (at least into English) is a conservative language variety: English-language translators tend to be slow in adopting innovative usages, preferring established variants (Redelinghuys, 2024).

References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015), Fitting linear mixed-effects models using lme4, *Journal of Statistical Software* 67(1), 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.

Biber, Douglas (1988), *Variation across Speech and Writing*, Cambridge: Cambridge University Press.

Herawati, Netti, Ameliana Wijayanti, and Agus Sutrisno (2024), The performance of ridge regression, LASSO, and elastic-net in controlling multicollinearity: A simulation and application, *Journal of Modern Applied Statistical Methods* 23. Available at: https://jmasm.com/index.php/jmasm/article/view/1258.

House, Juliane (2008), English as lingua franca in Europe today, in Guus Extra and Durk Gorter (eds), *Multilingual Europe: Facts and Policies*, Berlin: De Gruyter Mouton, 63–86. Available at: https://doi.org/10.1515/9783110208351.2.63.

Hu, Hai, and Sandra Kübler (2021), Investigating translated Chinese and its variants using machine learning, *Natural Language Engineering* 27(3), 339–372. Available at: https://doi.org/10.1017/S1351324920000182.

Kruger, Haidee, and Bertus van Rooy (2016), Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English, *English World-Wide* 37(1), 26–57. Available at: https://doi.org/10.1075/eww.37.1.02kru.

Kruger, Haidee, and Bertus van Rooy (2018), Register variation in written contact varieties of English: A multidimensional analysis, *English World-Wide* 39(2), 214–242. Available at: https://doi.org/10.1075/eww.00011.kru.

Lee, Bruce W., and Jason Hyung-Jong Lee (2023), LFTK: Handcrafted features in computational linguistics, *arXiv preprint* arXiv:2305.15878. Available at: http://arxiv.org/abs/2305.15878.

Lu, Xiaofei (2010), Automatic analysis of syntactic complexity in second language writing, *International Journal of Corpus Linguistics* 15(4), 474–496. Available at: https://doi.org/10.1075/ijcl.15.4.02lu.

Lu, Xiaofei (2017), Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment, *Language Testing* 34(4), 493–511. Available at: https://doi.org/10.1177/0265532217710675.

Nini, Andrea (2019), The Multi-Dimensional Analysis Tagger, in Tony Berber Sardinha and Marcio Veirano Pinto (eds), *Multi-Dimensional Analysis: Research Methods and Current Issues*, London: Bloomsbury Academic, 67–94.

Rabinovich, Ella, and Shuly Wintner (2015), Unsupervised identification of translationese, *Transactions of the Association for Computational Linguistics* 3, 419–432. Available at: https://doi.org/10.1162/tacl_a_00148.

Redelinghuys, Karien Reinette (2024), Language contact and change through translation in Afrikaans and South African English: A diachronic corpus-based study of genitive variation, in Bertus van Rooy and Haidee Kotze (eds), *Constraints on Language Variation and Change in Complex Multilingual Contact Settings*, Amsterdam: John Benjamins Publishing Company, 58–87.

Wickham, Hadley (2016), *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer-Verlag. Available at: https://ggplot2.tidyverse.org.

Xiao, Richard (2009), Multidimensional analysis and the study of world Englishes, *World* Englishes 28(4), 421–450. Available at: https://doi.org/10.1111/j.1467-971X.2009.01606.x.

Zou, Hui, and Trevor Hastie (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2), 301–320. Available at: https://doi.org/10.1111/j.1467-9868.2005.00503.x.

The impact of competitive European grant funding proposals and how to maximise it

Jolanta Šinkūnienė
(Vilnius University)

Work-In-Progress

Over the past decade there has been a growing surge of interest in competitive research funding discourse. Driven by the wish to unlock the success of the winning proposals of various national and international competitive grant programs, researchers look into rhetorical structures and lexico-grammatical features of texts representing competitive funding discourse, such as websites of projects (Lafuente-Millán, 2023; Mur-Dueñas, 2023), grant proposal summaries (Mėlinskas & Šinkūnienė, 2023) and grant applications (Millar et al., 2022; Millar et al., 2024). These studies reveal strategic and linguistic choices scholars make in order to

prove to peer reviewers and other members of the research community that the project they propose is novel, robust, original and ground-breaking.

Apart from the novelty and originality of research, one important element of competitive grant funding proposals is the description of the impact the proposed project is expected to bring. Impact could be achieved differently, it can address different needs, it can be immediate or it can be long-term, largely depending on the science field, and thus impact is not an easy category to predict, describe and evaluate. The focus of this paper is on the compulsory Impact section of the winning European networking projects funded under the European Cooperation in Science and Technology (COST) grants. Based on a self-compiled corpus of 50 Impact sections from projects representing various science fields, the paper employs qualitative and quantitative analysis to look into the discursive and linguistic choices used to describe the potential impact of the projects. The study also looks into the potential measures to maximise impact listed in these winning proposals. From traditional articles, summer schools and conference presentations to participating in hunting fairs and exhibitions to cross-media campaigns and live streams, the analysis reveals various ways to make outreach and maximise impact on industry, practitioners, media and civil organizations, academic audiences and the general public. The results of this paper could be useful to scholars interested in how to describe and communicate impact of their research, as well as to potential applicants to international research funding schemes.

References

Lafuente-Millán, Enrique (2023), European research project websites and related corporate websites: Patterns of evaluation and genre evolution, in J. Schmiedt, M. Bondi, O. Dontcheva-Navratilova and C. Pérez-Llantada (eds), (2023), *Patterns of Language Variation and Change in Academic Writing*, Special Issue of *Token: A Journal of English Linguistics* 16, 223–247.

Mėlinskas, Augustinas, and Jolanta Šinkūnienė (2023), The winner takes it all: Stance and engagement markers in successful project proposal abstracts funded by ERC, *Discourse and Interaction* 16(2), 98–123.

Millar, Neil, Bojan Batalo, and Brian Budgell (2022), Trends in the use of promotional language (hype) in abstracts of successful national institutes of health grant applications, 1985–2020, *JAMA Network Open 5*(8), e2228676.

Millar, Neil, Bryan Mathis, Bojan Batalo, and Brian Budgell (2024), Trends in the expression of epistemic stance in NIH research funding applications: 1985-2020. *Applied Linguistics* 45(4), 658–675.

Mur-Dueñas, Pillar (2023), Exploring researchers' professional digital discursive practices: A genre analysis of European research project websites, *Ibérica 45*, 79–107.

A corpus-based investigation of disciplinary differences and similarities of vague language in university lectures

Nicholas Smith and Amy Wang
(University of Leicester, Nottingham Trent University)

This paper investigates vague language in university lecture discourse, taking a corpus-based approach. It explores the similarities and differences across disciplines (i.e., Arts and Humanities, Social Sciences, Physical Sciences, and Life Sciences). In addition, the study examines the extent to which lectures differ from seminars, in terms of the frequencies and functions of vague language.

Investigations in English for Academic Purposes (e.g., Myers, 1996; Cutting, 2012, McCarthy, 2020) have established the strategic uses of vague language in functions such as guarded generalization and politeness. But in the context of classroom talk, vague language is under-researched. A rare example is Ruzaitė (2007), who explored two categories of vague language, quantifiers and approximators, in British

and American educational discourse. But there remain important research gaps on vague language in lectures and seminars specifically, and cross-disciplinary variation. Analysing vague language in seminar talk, Authors (2023) found marked differences between disciplinary groups. The present study builds on these foundations, focusing on lectures in British English universities.

Our research questions are:

1. What patterns of variation are found in frequencies and uses of vague language across the disciplinary groups in British university lectures?
2. What patterns of variation in vague language are found between lectures and seminars?

The primary data selected is the lecture component of the BASE corpus (British Academic Spoken English), consisting of 160 transcript files (40 per discipline group) and 1,155,974 words. For comparison purposes, we use the seminars in BASE, comprising 39 files distributed almost evenly across disciplinary groups (432,691 word tokens in total).

Drawing on a framework by Author (2010), we examined six categories of vague language: stance (e.g. *I mean*, *maybe*), quantifiers (e.g. *a bit of*, *some*), fuzziness (e.g. *sort of*, *more or less*), approximators (e.g. *about*, *approximately*), extenders (e.g. *things like that*, *and so on*), generic nouns (e.g. *things*, *stuff*). We conducted qualitative analysis of the pragmatic functions of these categories, notably politeness, persuasion, marking the speaker's stance, and effective use of language.

Regarding variation in lectures across disciplinary groups, our findings include that approximators and quantifiers increase in frequency as we proceed from Arts and Humanities through Social Sciences, Physical Sciences and Life Sciences**.** For fuzziness markers, the reverse pattern is found. Patterns for other vagueness categories are mixed, however. As for comparison between lecture and seminars, one finding is that whereas approximators, quantifiers and fuzziness markers follow a similar pattern in both discourse types, the other vagueness categories show less consistency. In our talk we suggest explanations for the observed quantitative and qualitative findings, and discuss their pedagogical implications.

References

Channell, Joanna (1994), *Vague Language*, Oxford: Oxford University Press.

Cheng, Winnie, and Martin Warren (2003), 'Indirectness, Inexplicitness and Vagueness Made Clearer', *Pragmatics*, 13(3), 381–400.

Cutting, Joan (2012), 'Vague Language in Conference Abstracts', *Journal of English for Academic Purposes*, 11(4), 283–293.

McCarthy, Michael (2019), 'Vague Language in Business and Academic Contexts', *Language Teaching*, 53(2), 203–214. https://doi.org/10.1017/S0261444819000100

Myers, Greg (1996), 'Strategic Vagueness in Academic Writing', in Eija Ventola and Anna Mauranen (eds), *Academic Writing: Intercultural and Textual Issues*, Amsterdam: John Benjamins, 3–17.

Walsh, Steve, and Anne O'Keeffe (2010), 'Investigating Higher Education Seminar Talk', *Novitas-ROYAL (Research on Youth and Language)*, 4(2), 141–158.

Ruzaitė, Jūratė (2007), *Vague Language in Educational Settings*, Frankfurt am Main: Peter Lang.

Thompson, Paul, and Hilary Nesi (2001), 'The British Academic Spoken English (BASE) Corpus Project', *Language Teaching Research*, 5(3), 263–264.

Modeling the 'evaluative attribution construction(s)' (EAC) and its development in Modern English

Veronika Stampfer
(FAU Erlangen-Nürnberg)

This study explores constructs such as the following:

(1) He still regards himself as a working farmer.ARCHER-1979-NEW
(2) We consider it almost one of the necessaries of life.ARCHER-1875-ADV
(3) He finds the bromide to be most suitable.ARCHER-1864-MED
(4) We no longer think of your form of worship as idolatrous.ARCHER-1966-SER

These examples share a common underlying Argument-Structure-Construction wherein an Attributor attributes an Attribute to an Attributee. In contrast to other Secondary Attribute Constructions (D'hoedt & Cuyckens 2017a) or what Quirk et al. (1985) call "Object Complement", this attribution does not alter the state of the Attributee like the "Resultative-Construction" with causative meaning but is a subjective evaluation which may not be factual. (Halliday 1967, 63. Compare: *He considers the couple husband and wife* doesn't mean the couple is actually married, whereas *He pronounced the couple husband and wife* does confer marital status.) This is why the construction, also known as "AGENT–ÆFFECTED–JUDGEMENT"-Construction (Herbst & Uhrig 2009), will be termed Evaluative-Attribution-Construction in this study, as the "evaluative" or "mental" (D'hoedt & Cuyckens 2017b) version of the Object Attribute Construction from Herbst's and Hoffmann's "Constructionist Approach to Syntactic Analysis" (CASA, Object Attribute Construction Attr:NP).

Despite semantic similarities, there is considerable formal variation – in other words considerably different fillers for the slots – which may also interact with finer-grained meaning variations of the construction. Aside from the verb slot, the construction differs in the Attribute-slot which can, e.g., be a Noun-Phrase as in (1) and (2) or an Adjective-Phrase (3 and 4). Moreover, the construction connects the Attributee with Attribute-slot with either the preposition *as* (1 and 4), without any filler as in (2) or with *to be* as in (3).

The first aim of this study is to discern whether EAC is best modeled as a single construction, multiple allostructions, or a family of constructions across abstraction levels. (cf. Cappelle 2006, Colleman 2011, Herbst & Huber 2022, Perek 2015, Goldberg & Jackendoff 2004) Second, slot-wise frequency analysis will trace its variation and development, examining verbal variants (e.g., *think*, *consider*) and variants of slots like Connector.

Preliminary findings suggest a decline in usage of this phenomenon; notably, the variants with Germanic verbs like *think* decline (1600-49: 313 hits vs. 1950-1999: 69 hits PMW-normalized), while the French loan words like *consider* as filler of the verb slot rise (1600-49: 10 hits vs. 1800-49: 70 hits, PMW-normalized). Aside from the constructional variants with high-frequency verbs, frequently cooccurring fillers are constructional candidates like the fillers *think* together with *fit* in the Attribute slot or *regard* together with the connector *as*. This data used in this study is taken from the ARCHER corpus (1600-1999) to examine slot variations and their development, contributing insights into the dynamic nature and structure of EAC(s) in Modern English.

References

ARCHER Consortium (2019), ARCHER 2019 – Beta Version 2*: A Representative Corpus of Historical English Registers*, 1990–1993/2002/2007/2010/2013/2016/2019. Originally compiled under the supervision of Douglas Biber and Edward Finegan; expanded by a consortium of universities. Example usages obtained under the terms of the ARCHER User Agreement.

Cappelle, Bert (2006), Particle placement and the case for 'allostructions', *Constructions* 1, 1–28.

Colleman, Timothy (2011), Ditransitive verbs and the ditransitive construction: A diachronic perspective, *Zeitschrift für Anglistik und Amerikanistik* 59(4), 387–410.

D'hoedt, Frauke and Hubert Cuyckens (2017), Language change in constructional networks: The development of the English secondary predicate construction, *Language Sciences* 59, 16–35.

D'hoedt, Frauke, and Hubert Cuyckens (2017), The development of the *as*-Secondary Predicate Construction: Constructionalization and internalization, *Language Sciences* 59, 16–35.

Goldberg, Adele, and Ray Jackendoff (2004), The English resultative as a family of construction, *Language* 80, 532–569.

Halliday, Michael A. K. (1967), Notes on transitivity and theme in English, Part 1, *Journal of Linguistics* 3, 37–81.

Herbst, Thomas, and Judith Huber (2022), Diachronic Construction Grammar – Introductory remarks to this special issue, *Zeitschrift für Anglistik und Amerikanistik* 70(3), 213–221.

Herbst, Thomas, and Peter Uhrig (2009), Erlangen Valency Patternbank: A corpus-based research tool for work on valency and argument structure constructions, Available at: www.patternbank.fau.de (last accessed: 3 May 2025).

Herbst, Thomas, Thomas Hoffmann, and Peter Uhrig (2018), CASA — A constructionist approach to syntactic analysis, Available at: https://constructicon.de/ (last accessed: 3 May 2025).

Perek, Florent (2015), *Argument structure in usage-based Construction Grammar*, Amsterdam and Philadelphia: Benjamins.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1986), *A comprehensive grammar of the English language*, 24th ed, London and New York: Longman.

Using Large Language Models for proficiency assessment of EFL learner writing

Bethany Stoddard, Lisa-Christine Altendorf, Robert Fuchs and Valentin Werner
(University of Bonn, University of Bamberg)

Essays are a typical task in (foreign) language teaching and standardized testing, and are traditionally evaluated by human raters, especially language educators. In addition, global proficiency assessments of learner texts, for instance following the Common European Framework of Reference for Languages (CEFR) categories (Council of Europe, 2001), are viewed as an important variable in second language acquisition and learner corpus research (Higgins et al., 2015; Wisniewski, 2017). Due to the high effort and costs of manual scoring, the prospect of Automatic Essay Scoring (AES) has attracted significant attention over the last half century, as methods for Natural Language Processing have advanced. AES tools have been in development since the 1960s (see Project Essay Grader, PEG Page, 1966; 1968). An updated version of the PEG is still in use today (Measurement Incorporated, 2024), in addition to a variety of other AES systems (see Lagakis & Demetriadis, 2021; Ramesh & Sanampudi, 2022; and Hussein et al., 2019 for overviews).

While extensive work has focused on the utility of AES systems for evaluating essays written by (primarily) English native speakers, the potential for AES systems to evaluate the proficiency of learners of English as a foreign language (EFL) has received far less attention. A few studies have used feature-based models as well as neural-network based models to predict CEFR levels from English learners' texts (Tack et al., 2017; Ballier et al., 2019; Caines & Buttery, 2020; Kerz et al., 2021; Gaillat, 2022), with classification accuracy ranging from 53% to 75%. Following the recent advancement of generative Artificial Intelligence (AI), a smaller number of studies have investigated whether Large Language Models (LLMs) can accurately rate learner essays on the CEFR scale. For example, Yancey et al. (2023) found that GPT-4 evaluated essays on the CEFR scale with performance comparable to other AES systems. Schmalz & Brutti (2022) had remarkable success using pre-trained BERT models, which reached a classification accuracy of 97.7%. However, the performance of other LLMs in comparison to human raters has not yet been empirically tested. Thus, this study investigates whether various LLMs can accurately evaluate texts written by English learners and categorize them by proficiency level.

We will test the capacity of LLMs to evaluate automatic essay scoring using the *ICLE500* dataset (Thwaites et al., 2024). This corpus consists of 500 texts by EFL learners with various L1s, which were rated on the CEFR scale by a team of trained judges. We will compare the ratings provided by six widely used LLMs (ChatGPT, Copilot, Claude, Gemini, Mistral, and BERT) against human ratings and with each other to investigate the models' accuracy. Findings will reveal the capabilities of generative AI for CEFR classification, demonstrating how LLMs can provide scalable, consistent estimations of learner proficiency. Automated CEFR classification lends itself to practical applications both in corpus linguistics, for the efficient annotation and categorization of learner data, as well as in language teaching and assessment, where accessible and automated feedback is essential.

References

Ballier, Nicolas, Thomas Gaillat, Andrew Simpkin, Bernardo Stearns, Manon Bouyé, and Manel Zarrouk (2019), A supervised learning model for the automatic assessment of language levels based on learner errors, in M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, and J. Schneider (eds), *Transforming Learning with Meaningful Technologies*, Springer, 308–320.

Caines, Andrew, and Paula Buttery (2020), REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, 5614–5623.

Council of Europe (2001), *Common European Framework of Reference for Languages: Learning, teaching, assessment,* Cambridge University Press.

Gaillat, Thomas (2022), Investigating the scope of textual metrics for learner level discrimination and learner analytics, in A. Leńko-Szymańska, and S. Götz (eds), *Complexity, accuracy and fluency in learner corpus research,* John Benjamins, 21–50.

Higgins, Derrick, Chaitanya Ramineni, and Klaus Zechner (2015), Learner corpora and automated scoring, in S. Granger, G. Gilquin, & F. Meunier (eds), *The Cambridge handbook of learner corpus research,* Cambridge University Press, 587–604.

Hussein, Mohamed Abdellatif, Hesham Hassan, and Mohammad Nassef (2019), Automated language essay scoring systems: A literature review, *PeerJ Computer Science* (5)*,* e208. https://doi.org/10.7717/peerj-cs.208

Measurement Incorporated (2024), Services. Retrieved November 18, 2024, from https://www.measurementinc.com/services

Kerz, Elma, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel (2021), Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs, in *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications,* Association for Computational Linguistics, 199–209.

Lagakis, Paraskevas, and Stavros Demetriadis (2021), Automated essay scoring: A review of the field, in M. S. Obaidat, S. Bilgen, K.-F. Hsiao, P. Nicopolitidis, S. Oktuğ, & Y. Guo (Eds.), *Proceedings of the 2021 IEEE International Conference on Computer, Information and Telecommunication Systems (CITS),* 1–6. https://doi.org/10.1109/CITS52676.2021.9618476

Page, Ellis B. (1966), The imminence of... Grading essays by computer, *The Phi Delta Kappan* 47(5), 238–243. Available at: http://www.jstor.org/stable/20371545

Page, Ellis B. (1968), The use of the computer in analyzing student essays. *International Review of Education* 14, 210–225. https://doi.org/10.1007/BF01419938

Ramesh, Dadi and Suresh Kumar Sanampudi (2022), An automated essay scoring systems: A systematic literature review, *Artificial Intelligence Review* 55, 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Schmalz, Veronica Juliana and Alessio Brutti (2022), Automatic assessment of English CEFR levels using BERT embeddings, in *Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-it 2021*, 293–299. Available at: https://books.openedition.org/aaccademia/10828?lang=en

Tack, Anaïs, Thomas François, Sophie Roekhaut, and Cédrick Fairon (2017), Human and automated CEFR-based grading of short answers, in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications,* Association for Computational Linguistics, 169–179.

Thwaites, Peter; Kollias, Charalambos; Kanistra, Voula, and Magali Paquot (2024), *ICLE500.* https://doi.org/10.14428/DVN/RIOSSC

Wisniewski, Katrin (2017), Empirical learner language and the levels of the Common European Framework of Reference, *Language Learning,* 67(S1), 232–253.

Yancey, Kevin P., Geoffrey Laflair, Anthony Verardi, and Jill Burstein (2023), Rating short L2 essays on the CEFR scale with GPT-4, in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*. Available at: https://aclanthology.org/2023.bea-1.49.pdf

Per syntax ad semantics: The benefits of enriching the parsed corpora of historical English with lexical semantic annotation

Tara Struik and Lena Kaltenbach
(University of Mannheim)

This paper describes the procedure to semi-automatically enrich the Penn Parsed corpora of historical English with semantic verb class annotation and its potential in application for studies on change at the syntax-semantics interface.

The annotation framework presented here is based on the seminal Manner/Result distinction by (Levin & Rappaport Hovav, 1995, 2019). This distinction in the aspectual class of dynamic verbs has long been known to be grammatically relevant (e.g., Fillmore, 1970). It determines which event schema is activated, and thus governs the constructions in which a verb may be used. For instance, Present-day English Manner verbs allow object deletion, as in (1a), whereas Result verbs do not, as in (1b).

(1) a. John swept (the floor).
b. John broke *(the vase).

The distinction is also cross-linguistically relevant, as it has been argued to lie at the core of Talmy's (2000) dichotomy between satellite-framing languages, which lexicalize the Result outside the core meaning of the verb, and verb-framing languages, which lexicalize the Result as part of the verb.

Grouping verbs according to their aspectual class would be a useful tool in the study of the diachrony of verb meaning and the morphosyntactic constructions related to this. As of yet, however, there is no exhaustive resource available for the inventory of English verbs which can be easily exploited in a quantitative way. Studies in this domain are typically conducted by relying on qualitative analysis of existing word lists and lexicographical sources, which is labour-intensive and may overlook quantitative differences between different items. The core meaning of a verb is relatively stable, however, and has been shown to change in predictable ways (Van Gelderen, 2018), making it feasible to annotate this feature in a corpus.

The annotation procedure follows the pipeline in Fig 1. The recent enrichment of the PPCME2 (Kroch, Taylor, & Santorini, 2000), PCMEP (Zimmermann, 2018) and PLAEME (Truswell, Alcorn, Donaldson, & Wallenberg, 2019) corpora by Trips and Percillier (2020) makes it possible to adopt a lemma-based approach to the annotation. We determine the aspectual class for each occurring lemma on the basis of lexicographical evidence from the MED (MED online, 2000/2018) and OED (Proffitt, 2015). We distinguish three classes based on the work by Levin & Rappaport Hovav: Stative (e.g. *know*), Manner (e.g. *sweep*) and Result (e.g. *break)* (MR annotation in Fig.1). We also incorporate recent insights into Result verbs from Beavers and Koontz-Garboden (2020) and Yu, Ausensi, and Smith (2023), who argue that Result verbs should be subdivided into two types: one which relates individuals to a change-of-state event (which we label Resultcos), such as *burn*, and one which relates an individual and a state (which we label Result-state), such as *redden* (RType annotation in Fig. 1). This produces a list of lemmas with verb class annotation which will be available in a standoff format. The list will be accompanied by a Python script which automatically enriches the lemmatized CorpusSearch files by adding the annotation as an attribute to each verb in the corpus.

The talk will present how this novel annotation may be used by showing how the influx of French vocabulary in the Middle and Early Modern Periods has changed the inventory of constructions expressing Result: French verbs contribute transitive Result-state verbs to the English lexicon, which originally mostly consisted of Result-cos verbs. Complimentary distribution effects in syntax confirm that verbs are copied together with their formal lexico-semantic features from French into English.

References

Beavers, John, and Andrew Koontz-Garboden (2020), *The Roots of Verbal Meaning*, Oxford: Oxford University Press.

Fillmore, Charles J. (1970), The grammar of hitting and breaking, in R. A. Jacobs, and P. S. Rosenbaum (eds), *Readings in English Transformational Grammar*, Waltham, MA: Ginn, 120–133.

Kroch, Anthony, Ann Taylor, and Beatrice Santorini (2000), *Penn Parsed Corpus of Middle English 2*, Philadelphia: Department of Linguistics, University of Pennsylvania.

Levin, Beth, and Malka Rappaport Hovav (1995), *Unaccusativity: At the Syntax–Lexical Semantics Interface*, Cambridge, MA: MIT Press.

Levin, Beth, and Malka Rappaport Hovav (2019), Lexicalization patterns, in R. Truswell (ed), *The Oxford Handbook of Event Structure*, Oxford: Oxford University Press, 394–425.

McSparran, Frances, et al. (eds) (2000/2018), *Middle English Dictionary Online*, Ann Arbor: University of Michigan Library.

Proffitt, Michael (2015), *Oxford English Dictionary* (3rd edn), Oxford: Oxford University Press.

Talmy, Leonard (2000), *Toward a Cognitive Semantics. Volume 2: Typology and Process in Concept Structuring*, Cambridge, MA: MIT Press.

Trips, Carola, and Michael Percillier (2020), Lemmatising verbs in Middle English corpora: The benefit of enriching the Penn-Helsinki parsed corpus of Middle English 2 (PPCME2), the parsed corpus of Middle English poetry (PCMEP), and a parsed linguistic atlas of early Middle English (PLAEME), in N. Calzolari et al. (eds), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille: European Language Resources Association, 7170–7178.

Truswell, Robert, Rhona Alcorn, James Donaldson, and Joel Wallenberg (2019), A Parsed Linguistic Atlas of Early Middle English, in *Historical Dialectology in the Digital Age*, Edinburgh: Edinburgh University Press, 17–38. Available at: https://doi.org/10.1515/9781474430555-007

Van Gelderen, Elly (2018), *The Diachrony of Verb Meaning: Aspect and Argument Structure*, New York: Routledge. Available at: https://doi.org/10.4324/9781315180335

Yu, Jianrong, Josep Ausensi, and Ryan Walter Smith (2023), States and changes-of-state in the semantics of result roots: Evidence from resultative constructions, *Natural Language & Linguistic Theory*. Available at: https://doi.org/10.1007/s11049-023-09570-9

Zimmermann, Richard (2018), *The Parsed Corpus of Middle English Poetry*, Manchester: University of Manchester.

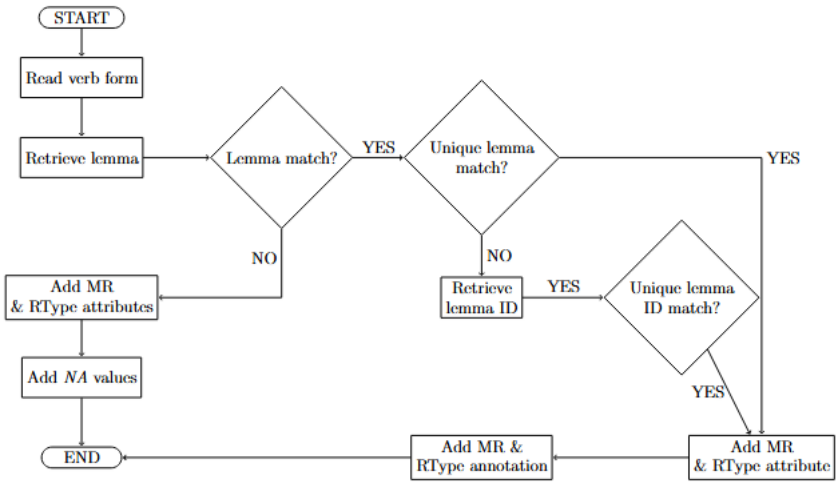


Figure 1: Annotation pipeline which adds Manner/Result (MR) and Result type (RType) annotation

LiTra – Linguistic Traces

Kjetil V. Thengs
(University of Stavanger)

Work-In-Progress

The proposed short paper is a presentation of a new project funded by an ERC Advanced Grant located at the University of Stavanger, Norway, for the period 2025-2029: LiTra – *Linguistic Traces: low-frequency forms as evidence of language and population history.*

The project aims to study low-frequency variation in language and explore its potential to reconstruct early language and population history, and will result in a unique text corpus of historical English documents, made searchable so that it can uncover very small patterns - linguistic traces - of earlier language contacts and interactions. Such traces may form a powerful means of reconstructing the past, especially when combined with the findings of genetics and archaeology, and help shed light on past linguistic areas as well as the survival of local and regional identities. The basic idea is that linguistic variation is everywhere, and we are surrounded by small patterns of rare words and structures that have yet to be studied systematically because they are so rare. At the same time, such patterns may carry information that is otherwise lost in the 'mainstream' of language change. The LiTra team are planning to make use of technological advances - including neural networks-based tools - to identify these kinds of 'micro-patterns' in historical texts.

The project team, which is also responsible for the *corpus of Middle English Local Documents* (MELD), is building on the existing MELD corpus to create a corpus of 4000 texts in total, collected by the team members from more than 70 different archives across the UK and beyond, and transcribed from original or facsimile using the transcription conventions of the *Middle English Scribal Texts program* (MEST) in Stavanger. All transcriptions are diplomatic, recording all orthographic information of the original texts. The texts included in the corpus are local documentary texts, which can be localized to a certain area and precisely dated based on explicit information in the texts, which makes them highly useful for the purposes of the LiTra project, but also for a range of other historical linguistics research. The historical text corpus will be a one-of-a-kind resource, and the plan is to annotate and lemmatize the corpus to make it searchable in various formats and for various purposes.

References

MEST – *The Middle English Scribal Texts programme.* https://www.uis.no/en/research/the-middle-english-scribal-texts-programme-mest

Stenroos, Merja, Kjetil V. Thengs, and Geir Bergstrøm (compilers) (2017), *A Corpus of Middle English Local Documents - MELD version 2017.1*. URL: http://www.uis.no/meld

Stenroos, Merja, and Kjetil V. Thengs (2020), *Records of Real People. Linguistic variation in Middle English local documents,* Amsterdam: John Benjamins

From past to present: Temporal shifts in rape discourse in the #MeToo era

Alessia Tranchese
(University of Portsmouth)

Introduction

This paper examines the diachronic evolution of collocational patterns associated with the term *rape* in a corpus of newspaper articles. In the context of the post-#MeToo era, it asks what changed – and what

remained consistent – in the portrayal of rape crimes in news coverage before and after the viral spread of the hashtag in 2017.

Data and Methods

The study uses a corpus of 89 million words, comprising articles about sexual violence published in the most widely read British newspapers between 2008 and 2019. Methodologically, it adapts the diachronic collocate analysis introduced by McEnery and Baker (2017) for application to a shorter timeframe, focusing on years rather than decades. This approach is used to investigate the connection between collocational and discursive change (or lack thereof), while also considering the broader social context and the cases that received the most media coverage during this twelve-year period. Additionally, the study employs the theory of lexical priming (Hoey, 2004) to demonstrate how the cumulative repetition of specific word combinations in rape-related news has driven observable language changes. This will be shown in particular through the study of the evolution of *alleged* as a collocate of *rape* over time and through the analysis of *rape* collocates relating to visible physical violence. The investigation is further enriched through the extraction of keywords from thematic sub-corpora built around specific search terms, employing an adapted concordance keywords method inspired by Taylor (2010) and Marchi (2010).

Key Results

The findings identify three enduring issues in the British media's coverage of rape crimes. First, the myth of "real rape" (Estrich, 1987) persists, reinforced by a strong association between rape and physically violent or fatal crimes, such as murder. Secondly, the level of mistrust towards victims has not decreased over time but has become subtler. This coincides with increased attention to rape cases involving high-profile perpetrators, particularly celebrities, athletes, and other powerful men. Third, media coverage continues to reflect a hierarchy of perpetrators, exhibiting disproportionate sympathy – what Manne (2017) terms "himpathy" – towards individuals perceived as vulnerable to false accusations, while harshly condemning "ideal perpetrators" (Boyle, 2019) for their "deviance."

In conclusion, this study argues that the widespread use of the #MeToo hashtag did not overlap with visible changes in rape discourse. Coverage of sexual violence remains fraught with problematic tropes, and most observed discursive changes occurred *prior* to the hashtag's viral spread. These changes are more closely linked to the rise of "celebrity culture" fostered by social media and the increasing dominance of online news platforms than to #MeToo itself.

References

Boyle, Karen (2019), *#MeToo, Weinstein and Feminism*, London: Palgrave Macmillan.

Estrich, Susan (1987), *Real Rape*, Cambridge, MA: Harvard University Press.

Hoey, Michael (2004), *Lexical Priming: A New Theory of Words and Language*, London: Routledge.

Manne, Kate (2017), *Down Girl: The Logic of Misogyny*, Oxford: Oxford University Press.

Marchi, Anna (2010), '"The Moral in the Story": A Diachronic Investigation of Lexicalised Morality in the UK Press', *Corpora*, 5(2), 161–189.

McEnery, Tony, and Helen Baker (2017), *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History*, London: Bloomsbury Academic.

Taylor, Charlotte (2010), 'Science in the News: A Diachronic Perspective', *Corpora*, 5(2), 221–250.

Understanding superficiality of LLMs for corpus linguistics: Opportunities and pitfalls with the example of L1 detection

Ahmet Uluslu and Gerold Schneider
(Zurich University)

Corpus linguistics can be portrayed as resting on three related pillars: a) corpus compilation, b) data extraction and annotation and c) data analysis. In recent years, large language models (LLMs) have demonstrated promising capabilities and new possibilities in many applications, including these pillars. We

focus on authorship analysis, where LLMs purportedly capture complex idiosyncratic features even in zero-shot settings (Huang et al, 2024). If that's true, a vast array of application in stylistics, literature and Digital Humanities is available, abstracting away from lexically centered research such as KWIC. However, growing concerns about LLM's faithfulness—whether they genuinely detect and explain linguistic patterns and generate human-like texts —necessitate deeper investigation (Agwar et al, 2024). Recent research demonstrates that LLMs can rely more on superficial textual elements, such as location and event descriptions, and generate texts following sequence information, rather than conducting thorough linguistic analysis (Uluslu and Schneider, 2024) and integrating enough contextual knowledge. The danger that they consistently take shortcuts is one of their main pitfalls, affecting their application to any corpus linguistic task.

Native language identification, the task of determining a writer's first language (L1) is based on their writing in a second language (L2), provides an ideal test case for investigating LLMs' linguistic capabilities, as it requires detecting subtle cross-linguistic patterns. This study addresses three research questions (RQ):

(1) To what extent do LLMs rely on topic- and sequence-based features instead of linguistic patterns in native language identification and generation tasks? This RQ addresses pillar a).

(2) What are the implications of topic-dependent analysis for high-stakes applications such as forensic linguistics? This RQ addresses pillar b). Does high automatic annotation performance translate to robust classification and features, or does it take corpus-dependent shortcuts? (Tognini-Bonellii 2001)

(3) How do different prompting strategies affect LLMs' analysis of linguistic features? We address pillar c) by systematically challenging the neural models.

We propose an evaluation framework using the TOEFL11 corpus (Blanchard et al. 2013). We systematically modify these essays to include critical topics (e.g., criminality, social conflicts) and varying degrees of emotional content, while carefully controlling for linguistic features. Superficial analyses can mislead forensic linguistics applications, where unreliable language analysis could have serious consequences in legal proceedings. Our methodology involves three components: (1) Creating controlled variations of the datasets using topic injection and sentiment modification techniques, (2) Developing a suite of prompting strategies ranging from zero-shot to few-shot settings with varying degrees of linguistic guidance, and (3) Implementing a comprehensive evaluation framework that measures both task performance and the model's attention to linguistic versus topic-based features. Through our investigation with state-of-the-art open-source LLMs, we examine how different prompting strategies can guide models towards more robust analysis, potentially mitigating their reliance on superficial features. Expected results will reveal the extent to which current LLMs rely on topic-based cues versus genuine linguistic signals, providing insights into their robustness for authorship analysis and further intricate stylistic tasks. We thus contribute to the development of more reliable and linguistically informed strategies for corpus stylistics applications.

References

Agarwal, Chirag, Sree Harsha Tanneru, and Himabindu Lakkaraju (2024), Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models, *arXiv preprint* arXiv:2402.04614.

Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow (2013), TOEFL11: A corpus of non-native English, *ETS Research Report Series* 2013(2), 1–15. Available at: https://doi.org/10.1002/j.2333-8504.2013.tb02331.x

Huang, Baixiang, Canyu Chen, and Kai Shu (2024), Can large language models identify authorship?, *arXiv preprint arXiv:2403.08213*.

Tognini-Bonelli, Elena (2001), *Corpus Linguistics at Work*, Amsterdam: John Benjamins.

Uluslu, Ahmet Yavuz, Gerold Schneider, and Can Yildizli (2024), Native language identification improves authorship attribution, in M. Abbas, and A. A. Freihat (eds), *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, Trento: Association for Computational Linguistics, 297-303. Available at: https://doi.org/10.18653/v1/2024.icnlsp-1.32

Tracing the dynamics of degree adverbs in Hong Kong English newspaper discourse: A study on genre development and nativization

Aditya Upadhyaya
(University of Würzburg)

Work-In-Progress

The interesting position of English in Hong Kong has sparked several studies on various features of Hong Kong English (HKE) (Peng and Setter, 2003; Joseph, 2004; Collins, 2009). While some linguists suggest that the L2 speakers of English in Hong Kong still adopt their linguistic norms from British or/and American English (Luke and Richards 1982: 51-52; Hyland 1997: 206), others view it as a localized variety (Bolton 2000: 274-276; Sung 2015: 266-267). However, the limited availability of diachronic corpora of Hong Kong English has resulted in a predominance of synchronic studies of this variety, which fail to explain a) the differences between changes in genre development and varietal development and, b) "to what extent a new postcolonial variety of English has diverged across time from the historical input variety" (Mukherjee and Schilk 2012: 191). The proposed study seeks to bridge these gaps by investigating the genre of newspaper writing in Hong Kong, using the diachronic corpus of Hong Kong English (DC-HKE), currently curated at University of Würzburg, Germany. Newspaper writing, as a genre, has the longest-standing tradition in Hong Kong, with English communication dating back to the establishment of the British colony in the mid-19th century until present day. While news writing constitutes a realm of professional language use with articles discreetly edited, a more comprehensive investigation of newspaper discourse may uncover subtle changes unlikely to be edited out. This study is particularly concerned with how stance is portrayed in HKE newspaper discourse using degree adverbs, which express the assertion of the speaker and the degree of the word which they modify. Stance-taking in HKE becomes an interesting premise, especially when considering the claims of Hong Kong media's self-censorship amidst political pressures and changing socio-political climate. To investigate the extent to which HKE can be said to have diverged from its input variety, the BLOB, LOB and FLOB databases of British English will be used. The study seeks to investigate the following facets:

1. **How has the frequency of degree adverbs in news discourse evolved over time? How far can HKE be said to have developed away from its input variety?**
2. **How do different factors like news theme, modified part of speech, semantic prosody, journalist's role affect the use of degree adverbs in HKE newspaper discourse?**

All chosen degree adverbs will be manually cleaned and coded for various independent variables such as part of speech modified, time period, semantic prosody, variety of English, journalist's role, etc. Next, multifactorial test (mixed effects regression) will be performed to determine "how much one can 'predict' what a response variable does depending on what one or more predictors do" (Gries 2021: 238). The study expects to uncover notable diachronic changes in the frequency and use of degree adverbs in HKE newspaper discourse. It is anticipated that the shift will reflect a gradual variation from British English, especially in response to local socio-political influences. The results of multifactorial analysis will further address research questions on genre development and structural nativization in HKE newspaper discourse.

References

Bolton, Kingsley (2002), The sociolinguistics of Hong Kong and the space for Hong Kong English, in Kingsley Bolton (ed), *Hong Kong English: Autonomy and Creativity*, Hong Kong: Hong Kong University Press, 29–55. https://hdl.handle.net/10356/96202

Collins, Peter (2009), Modals and quasi-modals in world Englishes, *World Englishes* 28(3), 281–292.

Gries, Stefan Th. (2021), *Statistics for Linguistics with R: A Practical Introduction*, Berlin/Boston: De Gruyter Mouton.

Hyland, Ken (1997), Language attitudes at the handover: Communication and identity in 1997 Hong Kong, *English World-Wide* 18(2), 191–210.

Joseph, John E. (2004), Case study 1: The new quasi-nation of Hong Kong, in *Language and Identity: National, Ethnic, Religious*, London: Palgrave Macmillan, 132–161.

Luke, Kang Kwong, and Jack C. Richards (1982), English in Hong Kong: Functions and status, *English World-Wide* 3(1) 47–64.

Mukherjee, Joybrato, and Marco Schilk (2012), Exploring variation and change in New Englishes: Looking into the International Corpus of English (ICE) and beyond, in Terttu Nevalainen and Elizabeth Closs Traugott (eds), *The Oxford Handbook of the History of English*, Oxford: Oxford University Press, 189–199. https://doi.org/10.1093/oxfordhb/9780199922765.013.0018

Peng, Long, and Jane Setter (2000), The emergence of systematicity in the English pronunciations of two Cantonese-speaking adults in Hong Kong, *English World-Wide* 21(1), 81–108. https://doi.org/10.1075/eww.21.1.05pen

Sung, Chit Cheung Matthew (2015), Hong Kong English: Linguistic and sociolinguistic perspectives, *Language and Linguistics Compass* 9(6), 256–270.

Collecting and analyzing student translations: A Lithuanian participation in the Multilingual Student Translation (MUST) project

Jurgita Vaičenonienė and Jonė Grigaliūnienė
(Vytautas Magnus University, Vilnius University)

Work-In-Progress

Although there have been prior initiatives to compile student translation corpora (e.g., PELCRA project (2001), Bowker's and Bennison's *Student Translation Archive* (2003), *MeLLange Project* (2007)), they were either limited by the pair of languages included, small-scale, inaccessible to wider researcher communities or no longer active. Therefore, initiated in 2016, the Multilingual Student Translation corpus (MUST) project, at present including partners from 20 countries, is one of the largest initiatives in the field of Translation Studies (Granger and Lefer, 2020). Within the particularly devised Hypal4MUST environment developed by Adam Obrusnik (Granger et al., 2019), project partners can collect and annotate parallel student translation corpora with rich metadata on participants, tasks, and source texts. The aim of this presentation is twofold. First, we would like to share the experience of the Lithuanian team participating in the project, to discuss the process and the results of collecting the Lithuanian student translation corpus MUST-LT, particularly its English-to-Lithuanian sub-corpus, and to demonstrate its applications in teaching (Kovalevskaitė et al., 2022). Second, we would like to touch upon the issue of a range of applications of the resulting corpus on the example of the text included as a task into the sub-corpus, METRO (173 words) (Baker 1992). It has also been included into the Hypal4MUST environment as a possible source text for potential collaborative and comparative cross-language research. Although, this advertising text is rich with different layers suitable for linguistic research, we will mostly focus on the qualitative analysis of student translations of selected informal and idiomatic phrasal verbs and collocations (*get up and go, after your own heart, don't hang around, get going, more than just a pretty face, just the ticket*). The data of 54 student translations who are training to be professional translators or have a background in English linguistics and are taking a course in translation was collected over the period of 2023-2024 along with student consent forms. The task was given as an in-class or a take-home activity; the students were allowed to use electronic dictionaries and other reference materials except for the automatic translation systems. Within the Hypal4MUST environment, the data was paragraph- and sentence-aligned (7095 Lithuanian words). The downloaded data will be anonymized and coded for informal and idiomatic phrasal verbs and collocations and their translation strategies. The accompanying discussion of student translation tendencies and their

effect on the overall meaning and texture of the advertising text might be valuable for pedagogical purposes in translator training as well as corpus-based translation studies in general.

References

Baker, Mona (1992), *In Other Words: A Coursebook on Translation,* Routledge.

Bowker, Lynne and Peter Bennison (2003), *Student Translation Archive: Design, Development and Application,* Routledge.

Granger, Sylviane and Marie-Aude Lefer (2020), The Multilingual Student Translation Corpus: A Resource for Translation Teaching and Research, *Lang Resources & Evaluation* 54, 1183–1199. https://doi.org/10.1007/s10579-020-09485-6

Granger, Sylviane, Marie-Aude Lefer, and Adam Obrusnik (2019), *Hypal4MUST: A Community-based Web Interface for Translation Teaching*, EUROCALL 2019, http://hdl.handle.net/2078.1/222250

Kovalevskaitė, Jolanta, Erika Rimkutė, and Jurgita Vaičenonienė (2022), *Lietuvių kalbos kolokacijos: vartojimas, mokymas(is) ir vertimas: mokomoji priemonė,* Kaunas: Vytauto Didžiojo universitetas. *https://doi.org/10.7220/9786094675249.*

*MeLLange Project (*2007), https://mellange.eila.univ-paris-diderot.fr/

Pezik, Piotr (2012), Towards the PELCRA Learner English corpus, *Corpus Data across Languages and Disciplines,* 33–42.

SCOTIA - towards the Scottish Corpus of original texts from immigrants to Aotearoa: First steps and findings

Sarah van Eyndhoven
(University of Canterbury)

Work-In-Progress

The development of spoken New Zealand English (NZE) has seen considerable analysis from a phonological perspective, thanks to the invaluable Origins of New Zealand English (ONZE) corpus, enabling complex theories of new dialect formation to arise (see e.g. Trudgill et al. 2000; Trudgill 2004; Britain 2005, 2008). Comparatively, much less attention has been directed to early written NZE (though see Hundt 2012; Hundt and Szmrecsanyi 2012 on the English, and Bonness 2017, 2019; Avila-Ledesma 2019 on an Irish migrant family), particularly regarding the Scottish setters arriving on New Zealand's shores, despite the fact that they consistently made up over 20% of European settlers to New Zealand (Bueltmann 2011). Recent literature has highlighted their strong and ongoing links to the 'homeland' and the dense social networks they established and maintained with other Scots upon arrival in New Zealand (McCarthy 2011). At the same time, the Scots were coming into contact with migrants from various British destinations as well as the indigenous Māori population (Wanhalla 2013).

Yet it is currently unclear whether these migrants adopted NZE and Māori lexis in their writing, or whether writing 'home' would in fact encourage greater Scots lexical use, and if this might be correlated with the writer's social aspirations and stylistic goals. This has been inhibited by the lack of a diachronic, text-searchable corpus of ego-documents covering the early European settlement of New Zealand. To address this current gap, the Scottish Corpus of Original Texts from Immigrants to Aotearoa (SCOTIA) is being created to explore this uncharted linguistic landscape. Specifically, correspondence written between 1848-1918 by first generation Scottish migrants, to family members back in Scotland, has been identified and photographed within various archive holdings and museums across New Zealand. Currently, the corpus contains 158 transcribed letters across 10 family collections, totalling circa 34,000 words. These letters are being tagged for region (where the letter was written from and presumably where the migrant settled), destination (where the letter was sent to), date, gender (of the sender), recipient (their relationship to the sender), mentions of location within the text, and instances of Scots, NZE and Māori lexis. These factors will

enable future quantitative sociolinguistic analysis into the role they might play in innovative or conservative lexical usage.

This work-in-progress-report details the process of sourcing this material, the use of Transkribus (Kahle et al. 2017) to automatically-transcribe the scanned material and tag it for linguistic and extralinguistic features, and the development of a custom-built platform in LaBB-CAT (Fromont and Hay 2008) to store and search the files. This followed by a qualitative analysis exploring the Scots, NZE and Māori lexis that has been found in these writings thus far and which semantic fields they are sourced from. Early investigations indicate that topography, cultural traditions and kinship terms are among the chief domains preserving Scots lexis, including *bairns* (children), *Hansel-Monday* (a holiday) and *brae* (hillside), whereas language related to the workplace encourages use of new dialect words, including *station* (sheep run), *drays* (carts) and *whare* (temporary hut). The plausible role of identity-projection, social mobility and maintenance of family connections influencing their use is demonstrated, highlighting the potential of corpora to explore the intersection between language, identity and dialect formation within a historical setting. Finally, future goals and next steps, including the exploration of morpho-syntactic and orthographic changes, and quantitative analysis of regional differences across areas of Scottish settlement within New Zealand, will be discussed.

References

Avila-Ledesma, Nancy E. (2019), "Believe my word dear father that you can't pick up money here as quick as the people at home thinks it": Exploring migration experiences in Irish emigrants' letters, *Corpus Pragmatics*, 3(2), 101–121.

Bonness, Dania Jovanna (2017), The Northern Subject Rule in the Irish diaspora: Subject-verb agreement among first-and second-generation emigrants to New Zealand, *English World-Wide*, 38(2), 125–152.

Bonness, Dania Jovanna (2019), '[S]eas may divide and oceans roll between but Friends is Friends whatever intervene'. Emigrant letters in New Zealand, in R. Hickey (ed), *Keeping in Touch: Emigrant Letters Across the English-Speaking World*, Amsterdam & Philadelphia: John Benjamins, 185–209.

Britain, David (2005), Where did New Zealand English come from? *Essex Research Reports in Linguistics*, University of Essex, 156–193.

Britain, David (2008), When is a change not a change? A case study on the dialect origins of New Zealand English, *Language Variation and Change*, 20(2), 187–223.

Bueltmann, Tanja (2011), *Scottish Ethnicity and the Making of New Zealand Society, 1850-1930,* Edinburgh: Edinburgh University Press.

Fromont, Robert and Jennifer Hay (2008), ONZE Miner: the development of a browser-based research tool, *Corpora*, 3(2), 173–193.

Hundt, Marianne (2012), Towards a corpus of early written New Zealand English – news from Erewhon? *Te Reo: Journal of the Linguistic Society of New Zealand*, 55, 51–74.

Hundt, Marianne and Benedikt Szmrecsanyi (2012), Animacy in early New Zealand English, *English World-Wide*, 33(3), 241–263.

Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger (2017), Transkribus - a service platform for transcription, recognition and retrieval of historical documents, *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 4, 19–24, IEEE.

McCarthy, Angela (2011), Scottish migrant ethnic identities in the British Empire since the nineteenth century, in J. M. MacKenzie and T. M. Devine (eds), *Scotland and the British Empire*, Oxford: Oxford University Press, 118–146.

Trudgill, Peter (1998), The chaos before the order: New Zealand English and the second stage of new-dialect formation, *Trends in Linguistics Studies and Monographs*, 114, 197–208.

Trudgill, Peter, Elizabeth Gordon, Lewis, and Margaret Maclagan (2000), Determinism in new-dialect formation and the genesis of New Zealand English, *Journal of Linguistics*, 36(2), 299–318.

Wanhalla, Angela (2013), *Matters of the heart: A history of interracial marriage in New Zealand,* Auckland: Auckland University Press.

Hardly sustainable or *a bit expensive*? Preadjectival intensifiers as markers of colloquialisation in parliamentary discourse

Turo Vartiainen and Turo Hiltunen
(University of Helsinki)

Colloquialisation, which refers to linguistic changes that reflect the relaxation of norms of written language towards increasing acceptance of spoken features (e.g., Mair 2024), has in recent research been observed to operate in a variety of written registers. These include both "agile" registers, such as newspaper prose, and more rigid ones, such as parliamentary discourse, which, although ultimately speech-based, represents highly regulated usage (Korhonen et al. 2023). In a recent study (Hiltunen & Vartiainen 2024), we found evidence of a colloquialisation trend in the *Hansard Corpus* (Alexander & Davies 2015), which comprises a comprehensive record of the speeches delivered in the British parliament in 1803–2005. However, our analysis also showed that colloquialisation progressed at different rates in the House of Commons and the House of Lords, which suggests that the texts/speeches produced in each House should potentially be treated as distinct sub-registers. As our previous study only considered a limited number of linguistic features, however, the generalizability of our findings is open for debate. In this paper, our goal is to shed more light on the relationship between changing linguistic norms (colloquialisation) and external factors (House) by investigating new data. Our case studies focus on intensifiers, more specifically downtoners (e.g., *rather*, *somewhat*, *a bit*; Quirk et al. 1985: 597-601). As intensifiers have variously been associated with such sociolinguistic correlates as (young) age, colloquial and nonstandard usage, the expression of emotions, and in-group membership (e.g., Tagliamonte 2008), their usage is predicted to correlate with colloquialisation. Examples (1) to (3) illustrate the kinds of items studied.

(1) I concede at once that this is to me a **rather** curious fact [...] (Commons, 1947)

(2) I am **somewhat** disappointed with the Bill because it does not go as far as one would wish. (Commons, 1990)

(3) I comment on Amendment No. 73 only to say that its spelling seems **a bit** shaky. (Lords, 1993)

To form a baseline for colloquialisation, we examined the register distribution of downtoners in Present-day English by consulting the *British National Corpus*. We provisionally classified downtoners into those associated with spoken language (e.g., *a bit*, *a little bit*) and those related to more formal, written registers (e.g., *somewhat*, *hardly*). By observing their frequency changes in the Hansard Corpus, we aim to answer three questions: i) has the style of British parliamentary discourse become more colloquial over time; ii) are there periods associated with particularly rapid change; and iii) does colloquialisation proceed at different rates in the House of Lords and the House of Commons.

Our case studies reveal long-term trends that can be interpreted in terms of a gradual colloquialisation of parliamentary discourse. Moreover, although the developments generally proceed in the same direction in both Houses, there are also substantial differences in the frequency of individual items across the Houses; this supports the idea that the texts from the two Houses represent distinct sub-registers. To complement our frequency-based study, we perform a collocational analysis of the data and consider various contextual and pragmatic factors that may explain our findings.

References

Alexander, Marc and Mark Davies (2015), Hansard Corpus 1803–2005. Available at: http://www.hansard-corpus.org

Hiltunen, Turo and Turo Vartiainen (2024), Corpus-pragmatic analysis of linguistic democratisation in the British Hansard. Comparing the two Houses, *Journal of Historical Pragmatics* 25(2), 245–273. Korhonen, Minna, Haidee Kotze, and Jukka Tyrkkö (eds) (2023), *Exploring Language and Society with Big Data: Parliamentary Discourse Across Time and Space*, John Benjamins.

Mair, Christian (2024), Colloquialisation: Twenty-five years on, *Journal of Historical Pragmatics* 25(2), 193–214.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985), *A Comprehensive Grammar of the English Language*, Longman.

Tagliamonte, Sali A. (2008), So different and pretty cool! Recycling intensifiers in Toronto, Canada, *English Language and Linguistics* 12(2), 361–394.

Using relative entropy to track disciplinary developments over time

Svetlana Vetchinnikova and Mikhail Zolotilin
(University of Helsinki)

How does a discipline evolve over time? How do individual members of the disciplinary community contribute to this development?

A common method to study conceptual change is by training word embedding models, but this has two drawbacks. First, these models require a large corpus (10 to 50 million words per period) for reliable results, which is often unattainable for smaller communities and individual authors. Second, the lack of transparency makes it hard to interpret results and connect them to linguistic processes of language change. This paper explores the use of Kullback-Leibler Divergence (KLD) to study the developments in the cognitive linguistic community and the trajectories of its individual members.

We collected a disciplinary corpus of articles published in *Cognitive Linguistics* between 1995–2024, totalling ca. 6.6M words, and an individual corpus of all articles published by a prominent cognitive linguist irrespective of the venue between 1977–2023, totalling ca. 1.4M words. Later, we plan to expand the disciplinary corpus to other journals and the individual corpora to other authors.

KLD measures the difference between two probability distributions. In studies of diachronic change, it can quantify the extent to which the distribution of words at time A diverges from that at time B. Following Degaetano-Ortlieb and Teich (2018), we aimed to identify periods of change and key contributing words using a data-driven approach. Instead of predefined time periods, we applied a one-year sliding window and a five-year period range. For example, in 2000, our KLD models compared articles from 2000–2004 to those from 1995–1999; in 2001, they compared articles from 2001–2005 to those from 1996–2000. To create the models, we lemmatized the corpora and set frequency (n=5) and dispersion (n=2) thresholds. In each year, we calculated the total KLD for all lemmas to inspect the overall trend and a KLD for each lemma to identify the most distinctive ones.

We found that in the communal corpus a major peak in KDL occurred around 2002–2004 and in the individual corpus around 1987 (Figure 1). Degaetano-Ortlieb and Teich (2018) suggest that a KLD peak indicates a period of lexical expansion and a KLD trough a period of lexical consolidation. In the context of small disciplinary and individual corpora, KLD peaks may be associated with the emergence of new topics and concepts.

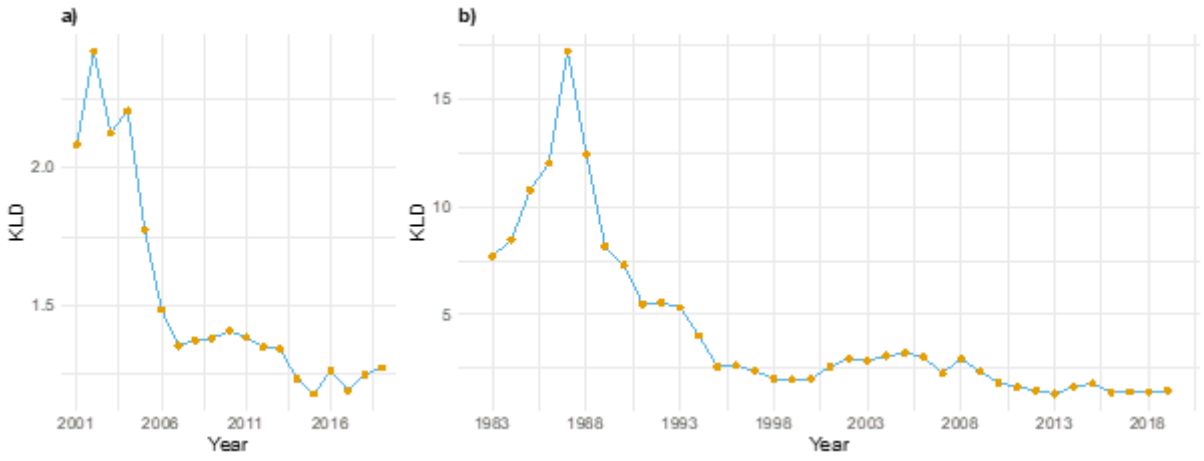


Figure 1. The change in the total KLD for lemmas in the communal (a) and individual (b) corpora over time.

To identify the most distinctive lemmas, we ranked them based on the range of their KLD values across all comparisons. This method highlighted lemmas such as *verb, dialogic, viewpoint, behaviour, children, think, construction, motion, gesture, narrative, metaphor,* and *vac* in the communal corpus. In the individual corpus, it highlighted lemmas such as *skill, spelling, sound, awareness, ability, keyword, vac, learning, explicit, segmentation, English, model, factor, implicit, rule,* and *knowledge*. We suggest that distinctive lemmas at the level of a community may reflect changes in disciplinary trends, while at the level of an individual scholar they may indicate changes in research interests.

Next, we will zoom in on the diachronic changes in the use of these lemmas by applying KLD to quantify the shifts in their collocate distributions over time (Tichý & Cvrček, 2024).

References

Degaetano-Ortlieb, Stefania and Elke Teich (2018), Using relative entropy for detection and analysis of periods of diachronic linguistic change, *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at COLING2018,* Santa Fe, NM, USA. ACL.

Tichý, Ondřej and Václav Cvrček (2024), *Detecting, analysing and visualizing semantic change: Collocational divergence in English and Czech,* Talk presented at the 45th ICAME Conference, Vigo (Spain), 18–22 June.

'You wanna flirt?' - 'Let's flirt!': A corpus-based analysis of flirting

Michelle Weckermann and Lena Scharrer
(University of Augsburg)

To say that flirting is an interesting linguistic and behavioural phenomenon would be nothing short of an understatement. It has no fixed form, yet one clear function: to signal sexual and/or romantic interest to (and in) an interlocutor (Speer, 2017; Kiesling, 2013). Most researchers agree on the fact that flirting is an ambiguous action that is difficult to define (Speer, 2017; Kiesling, 2013; Motschenbacher, 2020). Motschenbacher (2020) depicts it as a sexual speech act with the goal of expressing desire, while Kiesling (2013) and Speer (2017) emphasise its reciprocal and sequential nature.

The present study is of exploratory nature and aims to provide a comprehensive picture of flirting by approaching the phenomenon from different methodological angles. The study is corpus-based, drawing on naturally occurring data from the TV series *Love Is Blind (US),* where candidates get to know each other without visual information but with the ultimate goal of finding a partner for life. While flirting (and more generally, the negotiation of desire) has often been investigated in speed-dating encounters (see, e.g., Ranganath et al., 2009; Korobov, 2011; Stokoe, 2010), we consider *Love Is Blind* to be a suitable corpus, as it constitutes spontaneous spoken data, appears to be non-scripted, and is less routinised, time-restricted and thus also less systematic than speed-dating (see Stokoe, 2010 for a similar point).

The fact that flirting is difficult to pin down theoretically and empirically was reflected in the results. We took various methodological avenues to approach flirting, including metalinguistic references (e.g., 'Are you flirting with me / teasing me?'; search terms: *flirt*, teas**) and strings of humour (i.e., its perception through (co-)laughter, as this is one of the ways in which affiliation can be signalled; Gibbs et al., 2014). Additionally, flirting can consist of compliment sequences (Stokoe, 2010; Speer, 2017), can be prompted by self-praise (Speer, 2017), and can involve banter and teasing. Consider the example below:

Paul: I thought I told you last time.

Amber: No. Must have been your other girlfriend.
Paul: Must've been. [chuckles]
Amber: Eww!
Paul: I tease.
Amber: Literally nobody compares to me. [laughs]
Paul: Literally.
[*Love is Blind* S4, Ep.1]

The vehicles used for flirting – in the example above teasing (i.e. *Must've been, Must have been your other girlfriend*) and self-praise (i.e. *Literally nobody compares to me*) – furthermore seem to co-occur. While these individual linguistic phenomena have been studied (see, e.g., Dynel, 2008 for teasing; Dayter, 2018 for self-praise; Pomerantz, 1978 for compliments), these have to this day not been combined to try to paint an (if possible) all-encompassing picture of flirting. Motschenbacher (2020), for instance, approaches flirting via metalinguistic references, while Speer (2017) adds the reception of flirting, as well as how self-praise can function as a trigger for a flirting sequence.

We can thus say that flirting is a complex linguistic (and behavioural) phenomenon that relies on several vehicles to fulfil its function, including compliments, self-praise, banter, and teasing. Approaching flirting from these various angles can lead to a more complete picture of the phenomenon.

References

Dayter, Daria (2018), Self-praise online and offline: The hallmark speech act of social media, *Internet Pragmatics* 1(1), 184–203.

Dynel, Marta (2008), No aggression, only teasing: The pragmatics of teasing and banter, *Lodz Papers in Pragmatics* 4(2), 241–261.

Gibbs, Raymond W., Jr., Gregory A. Bryant, and Herbert L. Colston (2014), Where is the humor in verbal irony? *Humor: International Journal of Humor Research* 27(4), 575–595.

Kiesling, Scott F. (2013), Flirting and 'normative' sexualities, *Journal of Language and Sexuality* 2(1), 102–122.

Korobov, Neill (2011), Gendering desire in speed-dating interactions, *Discourse Studies* 13(4), 461–485.

Love Is Blind (US) (2024), Television series, Kinetic Content/Netflix.

Motschenbacher, Heiko (2020), Coming out – seducing – flirting: Shedding light on sexual speech acts. *Journal of Pragmatics* 170, 256–270.

Pomerantz, Anita (1978), Compliment responses: Notes on the co-operation of multiple constraints, in J. Schenkein (ed), *Studies in the Organization of Conversational Interaction*, Academic Press, 79–112.

Ranganath, Rajesh, Dan Jurafsky, and Dan McFarland (2009), It's not you, it's me: Detecting flirting and its misperception in speed-dates, in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 334–342.

Speer, Susan A. (2017), Flirting: A designedly ambiguous action? *Research on Language and Social Interaction* 50(2), 128–150.

Stokoe, Elizabeth (2010), Have you been married, or…?: Eliciting and accounting for relationship histories in speed-dating interaction, *Research on Language and Social Interaction* 43(3), 260–282.

Mapping the future of corpus preparation with ASR technology

Andreas Weilinghoff
(University of Koblenz)

The transcription of sound data is an essential yet time-consuming and labour-intensive part of almost all linguistic research projects. This is especially true for corpus linguistics, as only well-transcribed and carefully annotated corpora provide a reliable basis for subsequent analyses. As recent years have seen

great advancements in the field of Automatic Speech Recognition (ASR) (i.e. Watanabe et al. 2017, Radford et al. 2022), a central question is how the latest ASR models can be effectively used to speed up and enhance corpus transcription. More specifically, how do the latest ASR models perform in terms of accuracy and speed when compared to human transcribers?

This study will address this fundamental question. I will focus on how the end-to-end ASR system OpenAI Whisper (Radford et al. 2022) performs on sociolinguistic datasets. For this purpose, 120 sound files from different text categories of ICE Nigeria (Wunder et al. 2008) and ICE Scotland (Schützler et al. 2017) are re-transcribed with different Whisper models and the resulting transcriptions are then compared to the manual reference transcriptions via Word Error Rate (WER) metrics. To find out what significantly influences the performance of Whisper, the analysis applies linear mixed effects modelling of WER with the lme4 (Bates et al. 2015) and lmerTest (Kuznetsova et al. 2017) packages in R (R Core Team 2024). The individual files and speakers are treated as random factors. To compare Whisper's performance and speed with that of human transcribers, a subset of the data is re-transcribed by trained student assistants, with their working time closely monitored.

The findings show that Whisper performs well on both varieties. While the average WERs of 0.20 for ICE Scotland and 0.29 for ICE Nigeria are moderate, the ASR system effectively captures the main speech despite challenges such as strong accents and background noise. The inferential analysis further reveals significant influences of the *model* size, the *corpus*, the recording *quality*, the *text category*, the *speaker number* and the *gender* of speakers on the WER. A crucial issue of Whisper is that it automatically deletes hesitations, repetitions and interruptions which can pose a challenge for sociolinguistic data transcription. Nevertheless, the findings reveal that Whisper copes well with both inner circle and postcolonial outer circle varieties. Moreover, it can substantially accelerate corpus preparation compared to the transcription speed of human transcribers. On average, human transcription is 77% slower than that of the best Whisper models running on a regular laptop. At the same time, the transcriptions of most Whisper models are also more accurate than those of the student assistants. Based on the findings, I will discuss key opportunities and challenges of implementing Whisper for sociolinguistic data transcription and I will outline pathways for efficient transcription work with ASR technology and human transcribers.

References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015), Fitting linear mixed-effects models using lme4, *Journal of statistical software* 67, 1–48.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune HB Christensen (2017), lmerTest package: tests in linear mixed effects models, *Journal of statistical software* 82, 1–26.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2023), *Robust speech recognition via large-scale weak supervision,* Available at: https://arxiv.org/abs/2212.04356

R Core Team (2024), *R: A language and environment for statistical computing*. https://www.R-project.org/

Schützler, Ole, Ulrike Gut, and Robert Fuchs (2017), New perspectives on Scottish Standard English: Introducing the Scottish component of the International Corpus of English, *Perspectives on northern Englishes*, Berlin: De Gruyter Mouton, 273–302.

Watanabe, Shinji, Marc Delcroix, Florian Metze, and John R. Hershey (2017), New Era for Robust Speech Recognition, *Cham,* Switzerland: Springer International Publishing.

Wunder, Eva-Maria, Holger Voormann, and Ulrike Gut (2010), The ICE Nigeria corpus project: Creating an open, rich and accurate corpus, *ICAME Journal* 34 (1), 78–88.

Rhoticity in Singapore English: A corpus-based study of Singaporean media personalities

Cheryl Yeo
(Ludwig Maximilian University)

The variable realisation of coda /r/, henceforth referred to as (r), has been accounted for by the traditional phonological distinction of English varieties into rhotic and non-rhotic types (Wells 1982:218-220), and has been a hallmark area of study in sociolinguistic research (e.g. Labov 1966; Blaxter et al. 2019). Over the last three decades, the status of Singapore English (SgE) has been addressed in Tan and Gupta (1992); Poedjosoedarmo (2000); and Tan (2012) with a view to establishing whether this previously British-influenced variety is now being influenced by a more rhotic variety, i.e., American English. These previous studies have pointed to the media as a significant influencing factor in this change towards rhoticity, yet to date, there has been no known SgE study which has scrutinised speech in the media. The present study investigates a group that explicitly identifies as standing between global culture and Singaporean culture. The media personalities in this study position themselves as conduits of globalisation and modernity; however, they must still appeal to a Singaporean audience. Pertinent to this theme, are recent studies reporting how media personalities use a variety of linguistic styles (e.g. a globalised standard and a localised variety) to engage both international and local audiences (Lee 2014). Furthermore, media has become ever more embedded in daily life; this pervasiveness in today's contemporary world and styles of media personalities offer resources for everyday adoption and adaption in wider society (Coupland 2007:155; Bell and Gibson 2011:558).

This paper investigates the usage patterns of (r) in the speech of educated SgE speakers who are media personalities and seeks to answer the following research questions:

1. What are the observations on (r) in SgE speakers who are media personalities of varying ages, ethnicities and genders?
2. Which language-internal and language-external variables are significant predictors of the realised (r) in this variety of SgE?
3. Does the usage of (r) in the speech of Chinese, Malay, Indian, and Eurasian-Singaporean media personalities across various ages pattern the same way?
4. How would the findings of this study connect to the wider context of rhotacisation of other varieties of English, which have also been shown to display marginal rhoticity?

All speech data are conversational, spontaneous in nature, and extracted from publicly accessible media platforms, e.g. YouTube. To date, only speech data of Chinese and Malay speakers have been examined. To answer the first two research questions, the full sample of 3223 tokens is analysed by means of mixed-effects logistic regression using R (R Core Team 2021). Special attention with regard to how the factors – phonological context, preceding vowel, ethnicity, dominant language, and speech activity – influence (r) realisation is given. These were hypothesised to be significant predictors of rhoticity.

Separate regression analyses by ethnicity are conducted; outcomes from these analyses offer answers to the third research question of whether the (r) patterns of media personalities from different ethnic groups pattern similarly. A comparison of significant factors and constraint rankings are used to assess the similarity of these groups' systems. These are informative of emergent community grammars. The development of (r) in Chinese and Malay SgE speakers is starkly dissimilar; rhoticity patterns in the former appear to be advancing in a highly similar fashion with regard to rhotacisation of many other Englishes, whereas (r) realisation is strongly disfavoured in the speech of the latter, suggesting that Malay SgE speakers hardly participate in this phenomenon.

Along with quantitative analyses, qualitative analyses of certain influencers (e.g. Caitanya Tan and Kevin Tristan) in unique speech contexts, where style switching occurs, are implemented. These reveal that when speaking to an interlocutor, rather than to the masses, a localised non-rhotic variety of SgE is adopted. These analyses capture local indexical associations between variables and social meanings, thereby providing a nuanced picture of (r) usage.

The findings suggest that rhoticity in this variety of SgE is indeed present, but not uniform across all communities. Ethnicity, age, and certain linguistic factors play a vital role in this change toward rhoticity.

References

Bell, Allan, and Andy Gibson (2011), Staging language: An introduction to the sociolinguistics of performance, *Journal of Sociolinguistics* 15(5), 555–572.

Blaxter, Tam, Kate Beeching, Richard Coates, James Murphy, and Emily Robinson (2019), Each p[ɚ]son does it th[ɛː] way: Rhoticity variation and the community grammar, *Language Variation and Change* 31(1), 91–117.

Coupland, Nikolas (2007), *Style: Language Variation and Identity*, Cambridge: Cambridge University Press.

Labov, William (1966), The effect of social mobility on linguistic behaviour, S*ociological Inquiry* 36(2), 186–203.

Lee, Jerry Won (2014), Transnational linguistic landscapes and the transgression of metadiscursive regimes of language, *Critical Inquiry in Language Studies* 11(1), 50–74.

Poedjosoedarmo, Gloria (2000), The media as a model and source of innovation in the development of Singapore Standard English. In: Brown, A., D. Deterding, and L. E. Ling (eds), *The English Language in Singapore: Research on Pronunciation*, Singapore: Singapore Association for Applied Linguistics, 112–120.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. Available at: https://www.r-project.org/ (accessed 18 October 2024).

Tan, Chor Hiang, and Anthea Fraser Gupta (1992), Post-vocalic /r/ in Singapore English, *York Papers in Linguistics* 16, 139–152.

Tan, Ying Ying (2012), To r or not to r: Social correlates of /ɹ/ in Singapore English, *International Journal of the Sociology of Language* 218, 1–24.

Wells, John C. (1982), *Accents of English: Volume 1*, Cambridge: Cambridge University Press.

Organisers and Partners:

Vilnius University

Faculty of Philology

RESTORANAS

JOHN BENJAMINS PUBLISHING COMPANY

ACALA

Go Vilnius

VILNIAUS ROTUŠĖ

Rūta ANNO 1913