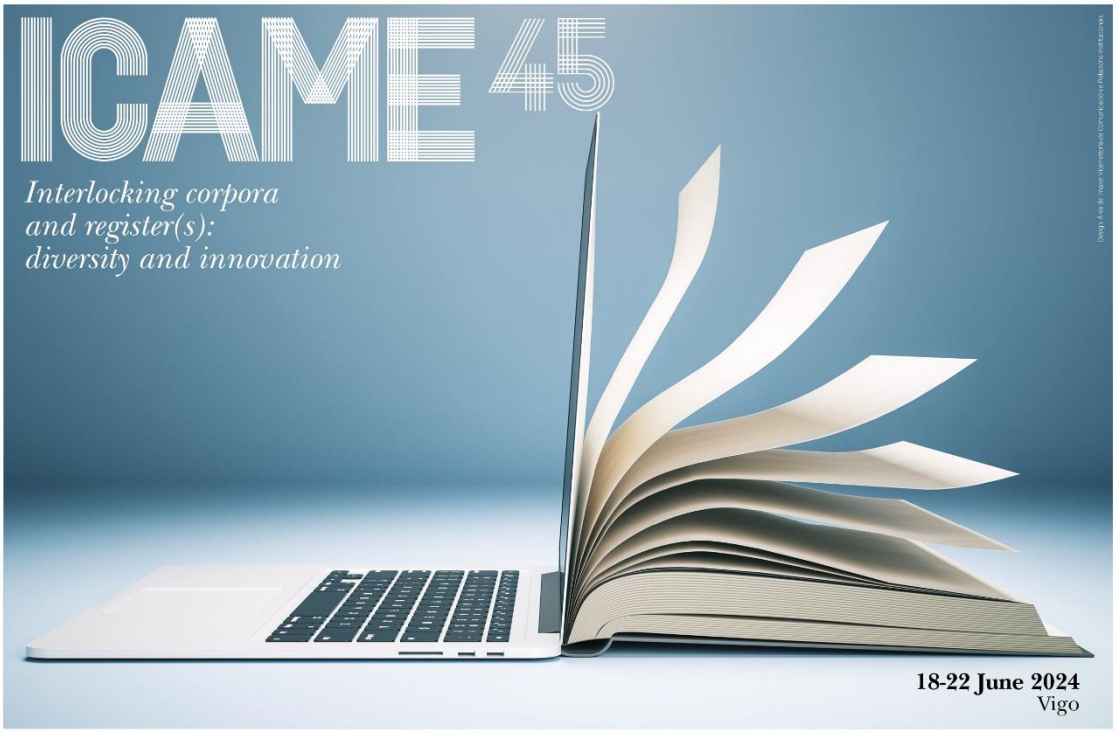
BOOK OF ABSTRACTS



Universidade Vigo

LVTC Research Group

lvtc

Contents

Plenary Speakers

Non-standardization: On the historical enregisterment of *ain't* in nineteenth-century American newspapers

Lieselotte Anderwald (University of Kiel)

The emergence of a "non-standard" register is quite obviously the flipside of standardization: if there is no standard, there can be no non-standard. If standardization is the suppression of optional variation (Milroy and Milroy 1999: 22), then a non-standard can be thought to emerge when optional variation is "relegated" to a register of non-standard, vernacular forms, or is reassigned non-standard status (i.e. stigmatized). Examples abound from the history of English that formerly optional variants persist (multiple negation, lack of adverb-marking <-*ly*>, non-standard verb forms, non-standard pronouns, different concord patterns, …), but at the price of heavy stigmatization. Perhaps the most stigmatized widely-used form today is *ain't*, a historically well-established negative contraction for all forms of present tense BE and HAVE (and, more recently, DO).

In this talk, I will try to show that "relegation to non-standard" is not an automatic by-product of codification or prescription, but a deliberate construction by interested parties, a process I will call "non-standardization". Taking *ain't* as my example, I trace the historical enregisterment of this negated verb in historical newspaper data (taken from the *AHN* database). The 19th century provides rich evidence of the increasing stigmatization of *ain't* in American English, through various text types that were common in newspapers at the time (local reports, anecdotes, political satire, letters, short fiction, even poetry). In these texts, characterological figures (in the sense of Agha 2007) are constructed and disseminated, often for the purpose of humour, that are characterized linguistically by their use of features like *ain't* (and, often, a host of other features). Laughing at these figures and the language they use consolidates them as stereotypes, and increasingly strongly links their language to their social characteristics, and their stigmatization.

The reconstruction of these stereotyped figures involves much qualitative, historical background work, and my talk thus also calls for a re-evaluation of qualitative, instead of solely quantitative, corpus work.

References

Agha, Asif. 2007. *Language and Social Relations*. Cambridge: Cambridge University Press.

AHN: America's Historical Newspapers: https://www.readex.com/products/americas-historical-newspapers

Milroy, James, and Lesley Milroy. 1999. *Authority in Language: Investigating Standard English*. 3rd edition. London & New York: Routledge. [First published 1985]

Flexible habits? Individual's responses to ongoing change

Hendrik De Smet (KU Leuven)

This study explores individuals' longitudinal behaviour with respect to ongoing change, asking how and to what extent individuals adjust to language change in their linguistic environment. To this end, data is used from the Hansard Corpus. The individuals in question are 68 prominent British politicians who were members of the House of Commons between 1920 and 2005. The changes are frequency shifts in 49 function words that have been selected bottom-up from the diachronic usage of the relevant speech community. A neural network has been trained to estimate the date of a given text based on the frequencies of the selected function words. Applied to individual usage data, the error of the neural network's estimates can be taken as a proxy to an individual's positioning with respect to the speech community. Overestimates suggest that an individual is on the whole ahead of the community, whereas underestimates indicate that an individual is on the whole behind on the community. Based on analysis of this data, it is argued that the major divide between patterns of individual longitudinal behaviour is not between progressiveness or conservatism. Instead, it is between speakers who generally adjust to population norms and track ongoing change, and speakers whose adjustments are only partial and accompanied by anomalous and apparently idiosyncratic developments. It is suggested that this divide reflects more basic psychological differences between speakers, whereby the feedback loop from exposure to mental representation and re-use is to a greater or lesser degree attuned to speakers' linguistic environment.

Manchester Voices: Creating a collection of Greater Manchester speech

Rob Drummond (Manchester Metropolitan University)

Manchester Voices is a large-scale sociolinguistics project that ran between May 2019 and January 2023. Its aim was to take a community-engaged approach in exploring the accents, dialects and identities of people living across the ten boroughs of Greater Manchester, a city-region in the Northwest of England.

Despite the challenges of a global pandemic on conducting community-engaged research, the project was able to collect spoken data from almost 500 people, creating a shareable corpus of voices from an important region of the UK. 200 of these people were visitors to our Accent Van – a mobile recording studio that we drove to community locations and events across the region. Others took part online in our Virtual Van, or in response to a call for people to 'Submit Your Voice' by completing an online spoken task. 350 people also took part in a perceptual study which explored folk-linguistic descriptions of, and feelings towards, language variety in the region.

In this talk I will discuss the benefits, practicalities and challenges of taking this kind of approach to sociolinguistics research and provide detail on some of the innovative methods we employed. I will then explore some of the data and findings, describing our various analytic tools and what they helped us to uncover. For example, the surprisingly distinct accent differences between the boroughs, the patterns of accent and dialect perception across the region, the topic-related style-shifting evident in archive recordings from the 1980s, and the personal insights people shared from their own experiences in relation to the way they speak and what it means to them.

Throughout the talk I will include the bits that went well, but also the bits that went less well, with the aim of encouraging us all to learn from the hugely enjoyable, but also difficult, imperfect and often messy experience of doing language research.

Cumulative knowledge building in corpus linguistics: Testing specific hypotheses about language and register

Tove Larsson (Northern Arizona University)

If we build knowledge in a systematic way, where subsequent studies are explicitly informed by previous studies, we can move the field's state-of-the-art forward more effectively. As corpus linguistics matures as a field, there are more and more research areas in which we may wish to build cumulative knowledge in this way. We can do so by (i) systematically synthesizing findings from previous research and interpreting new findings in relation to those, and (ii) formulating and testing increasingly specific hypotheses based on (i). For the former, the field arguably already has all the tools necessary; for the latter, however, our commonly-used approaches will only get us part of the way there.

For example, if we want to generalize to the population from which our corpus sample is drawn, we use inferential statistics (e.g., chi-square tests, various kinds of regression models). However, in their common usage, these methods do not enable us to explicitly build upon findings obtained from previous studies: Almost always, our null hypothesis is 'in the population, there is no difference or relation' and our alternative hypothesis is 'in the population, there is a non-zero difference or relation'. If our obtained *p*-value is below the α level (typically 0.05), we reject the null hypothesis and retain the alternative hypothesis. That is, given the traditional way we tend to use these techniques, we are stuck asking 'is there a difference/relation?' in an agnostic manner over and over, without being able to formally incorporate information obtained from prior studies about previously observed differences and relations (see Larsson, Biber, & Hancock, forthcoming).

In this talk, I will discuss benefits of cumulative knowledge building and how it can help us move the field forward. I will also report on findings from linguistic studies of register and grammatical complexity to show how specific hypotheses informed by previous findings can be tested through minimally sufficient statistical techniques that, rather than pulling us away unnecessarily, keep our focus firmly on the language data of interest.

References

Larsson, T., Biber, D. and Hancock, G. R. (Forthcoming). On the role of cumulative knowledge building and specific hypotheses: The case of grammatical complexity. *Corpora*, *19*(3).

Pre-conference Workshops

Workshop 1: Socio-Pragmatic Variation in Late Modern English

Convenors: Patricia Ronan and Christine Elsweiler (TU Dortmund, LMU Munich)

In recent decades it has been understood that pragmatics are not identical for all speakers and work in Variational Pragmatics has emerged which systematic differences between speakers of different national varieties, social classes, contexts, genres or registers are observed (e.g. Barron 2022, Ronan 2022). While historical varieties and diachronic changes are well researched overall, research on differences in earlier stages of different regional varieties mainly centres on formal linguistic categories, especially lexicon, syntax, or morphology. Research on pragmatic differences in earlier regional varieties of English is only in its infancy. Recently, Hudson (2023) has investigated the representation of language of the poor in fiction, Elsweiler (2021) and Elsweiler and Ronan (2023) have investigated evidence for pragmatic differences amongst regional varieties, and Elsweiler (2022) has researched gender-based variation in requestive behaviour. As we know now that different patterns of pragmatic variation are highly salient in contemporary varieties of language, we should research to a larger degree than has been done before what extent of pragmatic variation can be and should be expected for earlier varieties of English.

It is the aim of the proposed workshop on pragmatic variation in Late Modern English to help bridging this gap by bringing together researchers approaching the issue of historical pragmatics from different angles in order to reach synergy effects and to work towards a common framework of historical variational pragmatics.

References

Barron, Anne. 2022. Responses to thanks in Ireland, England and Canada: A variational pragmatic perspective. *Corpus Pragmatics* 6.2. 127-153.

Elsweiler, Christine. 2021. Divergence in two historical varieties: The use of modal auxiliaries in commissive and directive speech acts in Older Scots and Early Modern English letters. *Anglistik* 32:1.109-132.

Elsweiler, Christine. 2022. Gender variation in the requestive behaviour of Early Modern Scottish and English letter-writers? A study of private correspondence. *Journal of Historical Sociolinguistics* 8:1.55-88.

Elsweiler, Christine & Patricia Ronan. 2023. From *I am, with sincere regard, your most obedient servant* to *Yours sincerely*: The simplification of leavetaking formulae in 18th-century Scottish and Irish English letters. *ICAME Journal* 47.1. 1-17.

Hudson, Jane. 2023. Talking to peasants: language, place and class in British fiction 1800-1836. *English Language and Linguistics*. 27.3. 543-560.

Ronan, Patricia. 2022. Directives and politeness in SPICE Ireland. *Corpus Pragmatics*. 175-199.

Schneider, Klaus and Anne Barron (Eds.). 2008. *Variational Pragmatics*. Amsterdam and Philadelphia: John Benjamins.

I wonder if and *I would be grateful if*: The rise of new conventional indirect directives in Late Modern English

Laurel Brinton (University of British Columbia)

The most common (and most highly conventionalized) indirect directives in Present-Day English are *can/could/would you* DO X? (see Aijmer 1996: 147), addressing the preparatory condition on directives. These arise after 1900 in British English (Culpeper and Demmen 2011), becoming frequent in American English only in the second half of the twentieth century (Jucker 2020: 172–82). Late Modern English also sees the appearance of a large number of other (semi-)conventionalized indirect directives, including *I wonder if*, *Do you think?*, *Would you mind?*, *I was hoping*, *The best thing for you to do*, *I would/should be grateful/glad if*, *I would/should appreciate it if*, *Would you be so good/kind as to?*, *May/can I ask you to? You don't happen to?*, *Do you happen to?*, and others. Based on corpus evidence, these all appear for the first time in the middle of the nineteenth century. They function as "external modifiers" (Aijmer 1996: 170) and are classified by Leech (2014: 162–168) as belonging to the classes of "deliberative openings", "appreciative openings", "hedged performatives", and "happenstance indicators".

I wonder if (a deliberative opening) and *I would be grateful if* (an appreciative opening) are found to be the most frequent external modifiers of directives in the London-Lund corpus (see Aijmer 1996: 150). Neither usage is recognized in the OED. For Aijmer (1996: 153), *I would be grateful if* is "extremely tentative and formal", and *grateful* is often intensified. Leech (2014) sees *I wonder if* as occupying "the most indirect and most polite end of the pragmalinguistic politeness scale"; Huddleston and Pullum (2002: 941, 974) analyze it as a "doubly indirect directive", hence highly polite. *I wonder if* and also *I am wondering if/I wondered if* show a range of tentativeness (hedging) and are polite and formal (Aijmer 1996: 153, 163, 164). Wierzbicka (2006) proposes a scale of "scripts" which allow the "Anglo" speaker to avoid imposing or "putting pressure" on others. Beginning with imperatives and performatives (which place maximum pressure on the hearer), the language sees the rise of "whimperatives" (i.e. *can/could/will/would you* DO X?) and then a range of "suggestive" constructions (e.g. *you might like to/consider*, *would you be so good as?*, *would you mind?*, *perhaps you could*). These two stages place progressively less pressure on the hearer. The final stage beyond suggestions consists of *I was wondering if*; this serve to "avoid the impression that some pressure, however light, is being put on the addressee" (2006: 53). Although Wierzbicka implies that her scale is diachronic, she gives no evidence for this.

This paper will examine the origin and development of these forms and their variants in Late Modern English, probing their frequencies over time, their pragmatic and stylistic functions, their genre distribution, and their degree of conventionalization. The study will be primarily qualitative. A variety of Late Modern English corpora will be searched (CLMET3.1, CEAL, COHA, Founders) for the relevant directive search strings (*I/we wonder if you*, *I/we would be grateful if you*) and variants (*I was wondering/wondered if you*, *I wonder if it might be possible*, *I would be very grateful if you*, etc.) using either the in-built search programs or *AntConc*. Results will be manually post-edited.

The paper will speculate about the reasons for the exuberant rise of externally modified indirect directives in the nineteenth century, as part of the general shift towards non-imposition (negative) politeness in contemporary society (see Jucker 2011, 2012, 2020), the ethos of individualism (Culpeper and Demmen 2011), and democratization (Farrelly and Seoane 2012).

References

Aijmer, Karin. 1996. *Conversational routines in English: Convention and creativity.* London and New York: Longman

Anthony, Laurence. *AntConc*. Version 4.2.4. https://www.laurenceanthony.net/software/antconc/

Culpeper, Jonathan and Jane Demmen. 2011. Nineteenth-century English politeness: Negative politeness, conventional indirect requests and the rise of individual self. *Journal of Historical Pragmatics* 12(1–2). 49–81.
Farrelly, Michael and Elene Seoane. 2012. Democratization. In Terttu Nevalainen and Elizabeth Closs Traugott (eds.), *The Oxford handbook of the history of English*, 392–401. Oxford: Oxford University Press.
Jucker, Andreas H. 2011. Positive and negative face as descriptive categories in the history of English. *Journal of Historical Pragmatics* 12(1–2). 178–197.
Jucker, Andreas H. 2012. Changes in politeness cultures. In Terttu Nevalainen and Elizabeth Closs Traugott (eds.), *The Oxford handbook of the history of English*, 422–433. Oxford: Oxford University Press.
Jucker, Andreas H. 2020. *Politeness in the history of English: From the Middle Ages to the present day.* Cambridge: Cambridge University Press.
Leech, Geoffrey. 2014. *The pragmatics of politeness.* Oxford: Oxford University Press.
Oxford English dictionary. 2000–. 3rd ed. online. Michael Proffitt (ed.). www.oed.com
Wierzbicka, Anna. 2006. Anglo scripts against "putting pressure" on other people and their linguistic manifestation. In Cliff Goddard (ed.), *Ethnopragmatics: Understandind discourse in cultural context*, 31–63. Berlin: Mouton de Gruyter.

Corpora
Corpus of Early American Literature (CEAL). See Höglund, Mikko and Kaj Syrjänen. 2016. Corpus of Early American Literature. *ICAME Journal* 40(1). 17–38.
The Corpus of Late Modern English Texts, version 3.1 (CLMET3.1). Compiled by Hendrik De Smet, Susanne Flach, Hans-Jürgen Diller, and Jukka Tyrkkö. See https://fedora.clarin-d.uni-saarland.de/clmet/clmet.html
Davies, Mark. 2010. *The Corpus of Historical American English* (COHA). Available online at https://www.english-corpora.org/coha/
Founders Online. US National Archives and Records Administration Commission (NHPRCP) in partnership with the University of Virginia. https://founders.archives.gov

---

Fom *pray* to *please*: Sociopragmatic patterns in the Old Bailey Corpus (1720-1913)

Claudia Claridge (University of Augsburg)

Late Modern English (LModE), in particular the nineteenth century, is the crucial period for the switch from *pray* to *please* as the typical politeness marker in requests. Jucker (2020: 171), for example, shows the steep rise of *please* in the *Corpus of Historical American English* (COHA). *Pray* (originally *I pray you*) shows speaker-orientation and fairly bluntly marks the requesting intention. *Please* (originally *if you please*), in contrast, pays attention to the face needs of the hearer, has a tentative nature, and works within the negative-politeness / non-imposition approach that is typical of LModE (Culpeper & Demmen 2011).

Despite its restriction to the formal courtroom situation, the *Old Bailey Corpus* (OBC) offers an ideal opportunity to investigate this change. It contains speech-based interactive material from the courtroom and quotes from everyday interactions, all produced by socially diverse speakers within a mostly asymmetric power situation. Formally, *pray* appears only in its most pragmaticalized form, while *please* proceeds through verbal uses like *an't please you, if you please, if it please you, please you* (etc.) to the one-word marker *please*. The first unambigous instance of the latter appears in 1839, only shortly before the last occurrence (of only 16 in the 19th century) of *pray* in 1851. The paper will document the frequency distribution of the forms across time and across speaker groups. *Pray* is preferred by higher social classes (marginally)

and especially by judges and lawyers, i.e. speakers with institutional power in the courtroom, who therefore do not have to consider hearer face needs. *Please* is preferred by lower-class speakers, as well as lay speakers, i.e. victims, witnesses, and especially defendants. These speakers usually address superiors in the courtroom and for defendants a less confrontative requesting behaviour might have been appropriate. Beyond the courtroom the instances quoted by speakers may further offer insights into changing conventions.

References

Culpeper, Jonathan, and Jane Demmen. 2011. Nineteenth-century English politeness: Negative politeness, conventional indirect requests and the rise of the individual self. *Journal of Historical Pragmatics* 12.1–2: 49–81.

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

Stability of pragmatic markers: The case of *sorry* in organizational emails from the Clinton Email Corpus

Rachele De Felice (The Open University)

The focus of this study is the pragmatic variation across different writers and situations in their use of *sorry* (both as a stand-alone word and in phrases such as *I am sorry*). This presentation looks at the use of the word *sorry* in the Clinton Email Corpus. This is a collection of over 33,000 emails dating from Hillary Clinton's tenure as US Secretary of State, which has been released to the public following an investigation by the FBI. The dataset is an invaluable source of insights into organizational communication, not just because of its size, but also because it contains a wide range of senders and recipients whose identity, and therefore role within the organization, is publicly known. This in turn allows us to easily incorporate variables such as status and gender into any linguistic analysis.

All 545 occurrences of the term in the corpus are examined, using both quantitative and qualitative methods. Each occurrence is manually annotated for the type of action being apologised for, such as misunderstanding, unfortunate event, minor mishaps, and so on. Results show that the majority of instances of *sorry* refer to minor incidents such as misunderstandings, missed calls, and problems with emails, as well as being used for events ranging from typos to sad news regarding someone's poor health. In other words, it is more of a rhetorical or discursive device than a true pragmatic act. This holds regardless of the status of sender and recipient.

I argue that, beyond providing useful examples of typical email phraseology, the real value of corpus-based studies such as this one lies in unlocking the expected behaviour norms of the organization by showing us what its members deem necessary to be excused or apologised for (an overlong email, repeated missed calls, a lost schedule). The quantitative corpus investigation is the way into a broader qualitative interpretation of the context it reflects.

The organisation of macro-requests in early eighteenth-century Scottish and English letters

Christine Elsweiler (LMU Munich)

Previous research has shown that requests in early modern Scottish and English non-private letters manifest pragmatic variation regarding the use of internal modification through modal auxiliaries (Elsweiler 2021, 2023). In this study, I aim to complement this micro-level perspective on pragmatic variation in the realisation of requests by taking a macro-level approach. Requests in letters typically do not occur in isolation but are frequently organised into hierarchically structured speech act sequences (see van Dijk 1980: 184), i.e., macro-requests, which comprise different individual speech acts supporting the core request as pre- and post-moves, as is illustrated in the following example from an early 18th-century Scottish letter:

(1) Tho I haue not yet ben soe happy as to recue the anceuer of mine to you, yet I cannot but giue you this trouble,
to beg not only your aduis
which I find is the best to me of any,
but your assistance to uptaine what I desier, which is an act of councell in my faour,
(Francis Herbert, Countess Dowager of Seaforth to unspecified addressee, 1701)

In this macro-request, the discontinuous core request in (1) (in boldface), through which the writer seeks the addressee's advice and assistance, is supported by a preceding apology for troubling them with her petition. This is combined with a compliment for the addressee (*which I find is the best to me of any*), a convivial move to gain their favour.

Addressee-oriented expressive speech acts such as the compliment in (1) as well as, e.g., thanking and congratulating have been found to be central to 18th-century polite linguistic behaviour (Taavitsainen and Jucker 2010: 159), whose ideal is encapsulated in the phrase "the art of pleasing in conversation" (Jucker 2020: 120). Since in the 18th century, letters were conceptualised as written conversation (Klein 1993: 35), in a previous study (Elsweiler under review) I explored whether addressee-oriented expressive speech acts also feature prominently as supportive moves in Scottish letters from the first half of the 18th century. Specifically, I investigated whether their use saw an increase in Scottish letters written between 1570 and 1750. However, no such increase was discernible. Instead, letter-writers manifested a preference for writer-oriented speech acts such as commitments and apologies as supportive moves.

The present study will build on these findings to approach macro-requests from a variational pragmatic perspective by comparing the organisation of longer speech act sequences in Scottish and English letters written between 1700 and 1750. Since most research on 18th century politeness has been conducted for English data (e.g., Taavitsainen and Jucker 2010), this study aims to explore if addressee-oriented expressive speech acts are more centrally represented as supportive moves in macro-requests in English letters than in the Scottish letters.

The correspondence data for this study are drawn from the 18th century sub-section of *ScotsCorr* and from *CEECE*. The comparison is based on 80 Scottish and English letters, respectively, from which the macro-requests are manually retrieved and categorised according to a classification scheme developed for the analysis of macro-speech acts (see Elsweiler 2024).

References

CEECE = *Corpus of Early English Correspondence Extension*. 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 Modern Languages, University of Helsinki.

Elsweiler, Christine. 2021. Divergence in Two Historical Varieties: The Use of Modal Auxiliaries in Commissive and Directive Speech Acts in Older Scots and Early Modern English Letters. Daniela Kolbe-Hannah and Ilse Wischer (eds.). Focus on English Linguistics: Varieties Meet Histories. *Anglistik* 32(1): 115–138.

Elsweiler, Christine. 2023. Modal *May* in Requests: A Comparison of Regional Pragmatic Variation in Early Modern Scottish and English Correspondence. *Journal of Historical Pragmatics*. Online First. 1–37.

Elsweiler, Christine. 2024. Towards a Speech Act Annotation Scheme for 18th-century Scottish Letters. Christine Elsweiler (ed.). *The Languages of Scotland and Ulster from a Global Perspective – Past and Present: Selected Papers from the 13th Triennial Forum for Research on the Languages of Scotland and Ulster, Munich 2021*. Aberdeen: FRLSU. 246–279.

Elsweiler, Christine. Under review. The Conventional Organisation of Request Sequences in Scottish Letters (1570–1750). *Language and Literature*. Special issue on Diachronicity in Literary Studies and Linguistics. Ed. by Monika Fludernik and Olga Timofeeva.

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

Klein, Lawrence. 1993. Politeness as Linguistic Ideology in Late Seventeenth and Early Eighteenth-century England. Stein, Dieter and Ingrid Tieken-Boon van Ostade (eds). *Towards a Standard English: 1600–1800*. Berlin and New York: Mouton de Gruyter. 31–50.

ScotsCorr = The Helsinki Corpus of Scottish Correspondence 1540–1750. 2017. Ed. Anneli Meurman-Solin. Helsinki: University of Helsinki. http://urn.fi/urn:nbn:fi:lb-201411071

Taavitsainen, Irma and Andreas Jucker. 2010. Expressive Speech Acts and Politeness in Eighteenth-Century English. Raymond Hickey (ed.). *Eighteenth-Century English: Ideology and Change*. Leiden: Cambridge University Press. 159–181.

Van Dijk, Teun. 1980. *Macrostructures*: *An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Hillsdale, N.J.: L Erlbaum Associates.

The where, when, and how of speech: Variation and change in the direct speech representation formula in Late Modern English

Peter J. Grund (Yale University)

Much research has been devoted to reconstructing the spoken language of the past (e.g., Culpeper and Kytö 2010). As we only have access to written representations for most periods of the history of English, such reconstruction inevitably has to contend with the limitations of the written medium to capture facets of speech. While the writing strategies that language users employ to represent speech in historical periods are often seen as mere "filters" to be removed in the search for the underlying spoken language (e.g., Schneider 2013), they are in themselves important objects of study: they reveal significant variation and change over time as language users experiment with and hone linguistic tools and their functions to represent the voices of others (Grund 2023; forthcoming).

This paper focuses on one such aspect of speech representation that has received little attention in historical linguistic research: what I here call the "direct speech representation formula," as in the italicized formulation in "'What is it, Darwin? speak up!' *said Wharton, dropping at once into the colloquial tone, and stooping forward to listen*." (CLMET 3.0; 1894, Ward, *Marcella*). This "formula" involves the indication of the speaker and the speech representation verb as well as any concomitant description of time, place, manner, concurrent action, etc. of the speech represented. Some research has investigated aspects such as the nature of the verb and the order of the subject and verb (e.g., Ruano San Segundo 2016; Cichosz 2019), and some scholars have pointed to the presence of various features together with these speech representation expressions, usually in passing (e.g., Oostdijk 1990; Busse 2020; Hauff 2021).

Drawing data from the narrative fiction texts in CLMET 3.0, which covers the period 1710–1920, I bring these aspects together, charting the nature of the "formula" and the frequency and function of its various components. I consider change over time and across different authors, as well as a range of potentially influential.

Overall, this paper provides a systematic picture of the variation and change in the appearance of the direct speech representation formula and the strategic use of components of the formula for communicative and pragmatic purposes. As such, it contributes to the study of variation and change in Late Modern English influenced by pragmatic and communicative needs, which has received relatively little attention (cf. Lewis 2012). The paper also illustrates how literary texts offer important data for the study of speech representation in the history of English.

References

Primary Source

CLMET3.0 = Corpus of Late Modern English Texts, version 3.0. Created by Hendrik De Smet, Hans-Jürgen Diller, and Jukka Tyrkkö. https://perswww.kuleuven.be/~u0044428/clmet3_0.htm

Secondary Sources

Busse, Beatrix. 2020. *Speech, Writing and Thought Presentation in 19th-Century Narrative Fiction: A Corpus-Assisted Approach*. Oxford: Oxford University Press.

Cichosz, Anna. 2019. Parenthetical Reporting Clauses in the History of English: The Development of Quotative Inversion. *English Language and Linguistics* 23(1): 183–214.

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

Grund, Peter J. 2023. Disgusting, Obscene and Aggravating Language: Speech Descriptors and the Sociopragmatic Evaluation of Speech in the Old Bailey Corpus. *English Language and Linguistics* 27(3): 517–541.

Grund, Peter J. Forthcoming. Speech Representation in the History of English. In Joan C. Beal (ed.), *The Cambridge History of the English Language*, vol. 3, *Transmission, Change, and Ideology*. Cambridge: Cambridge University Press.

Hauff, Christoph Anton Xaver. 2021. *Verbs of Speaking and the Linguistic Expression of Communication in the History of English*. Berlin: Peter Lang.

Lewis, Diana M. 2012. Ch. 56: Late Modern English: Pragmatics and Discourse. In *English Historical Linguistics: An International Handbook*, Alexander Bergs and Laurel J. Brinton (eds), vol. 1, 901–915. Berlin: Mouton de Gruyter.

Oostdijk, Nelleke. 1990. The Language of Dialogue in Fiction. *Literary and Linguistic Computing* 5(3): 235–241.

Ruano San Segundo, Pablo. 2016. A Corpus-Stylistic Approach to Dickens' Use of Speech Verbs: Beyond Mere Reporting. *Language and Literature* 25(2): 113–129.

Schneider, Edgar W. 2013. Investigating Historical Variation and Change in Written Documents: New Perspectives. In J. K. Chambers and Natalie Schilling (eds.), *The Handbook of Language Variation and Change*, 2nd ed., 57–81. Oxford: Blackwell.

One of these things must be done – Move structure variation in Late Modern English threatening letters

Theresa Neumaier (TU Dortmund)

Threatening letters can take a huge variety of different forms: they might consist of a single sentence scribbled on a piece of paper but can also involve several pages of text. Nevertheless, they constitute a highly recognisable albeit illicit genre (Bojsen-Møller et al. 2020) with a clear underlying social function – the speaker declares their intention to carry out a harmful action against the recipient of the threat. While this main function seems to be shared by all types of threatening letters, additional functions, such as venting anger or manipulating the target into doing a specific action, can be identified for sub-categories of threats. As has been found in previous research, these more specific functions seem to influence the structure of the letter. The structure of extortion letters, for instance, can be compared to that of business letters, as it involves similar functional moves, such as a demand and declaration of consequences, an allocation of responsibility to the target, or a statement of sincerity (Busch 2006; Bredthauer 2020). However, research has mainly focused on present-day data so far, and genre conventions can change over time. Hence, is not yet clear whether these findings also hold for previous periods of English.

In this paper, I analyse variation in the structure of rhetorical moves in a small corpus of 100 threatening letters which I extracted from Old Bailey trial records as well as postings in the *The Gazette*, which regularly printed anonymous threatening letters with the promise of royal pardon to anybody willing to provide information about the letter writers. All of the letters were written between the 18th and early 20th century. The dataset is balanced with respect to whether the letter contains explicit conditions which the recipient is to fulfil to prevent the threat from being carried out or not. Hence, half of the data consist of extortion letters; the other half can be categorised as retaliative letters which lack a conditional element.

I show that the letters vary considerably with respect to the type and number of structural elements which are realised. As expected, some of this variation can be attributed to the overarching function of the letter, and hence appears to be linked to the writer's intention to manipulate the addressee or express a desire for revenge. However, other, more specific, factors also seem to play a role. These include the social circumstances in which the letter was written; for instance, when a writer is explicitly situating their text within the context of larger social grievances, such as agricultural protests of the time. This additional move widens the intended audience of the letter beyond the actual recipient themselves, which makes this particular type of threatening communication more similar to genres related to political protest. The study shows that threatening letters in the Late Modern English period form a clearly recognisable but nevertheless multi-faceted genre whose conventions are creatively negotiated by its users in their social contexts.

References

Bojsen-Møller, Marie, Sune Auken, Amy J. Devitt and Tanya Karoli Christensen. 2020. Illicit genres: The case of threatening communications. *Sakprosa* 12(1), 1-53.

Bredthauer, Stefanie. 2020. Erpresser- und Drohbriefe. In Marie I. Matthews-Schlinzig, Jörg Schuster and Jochen Strobel (eds.), *Handbuch Brief: Von der Frühen Neuzeit bis zur Gegenwart.* Berlin, Boston: De Gruyter, 594-600.

Busch, Albert. 2006. Textsorte Erpresserschreiben. In Sigurd Wichter and Albert Busch (eds.), *Wissenstransfer: Erfolgskontrolle und Rückmeldungen aus der Praxis.* Frankfurt a. M.: Peter Lang, 51-65.

Social variation in pragmatic markers in LModE Irish English letters: What can statistical approaches to greeting and leavetaking formulae show us?

Patricia Ronan (TU Dortmund)

The present study asks whether we can find indications of pragmatic variation based on status differences in Late Modern Irish English letters. While both social context and relative status differences between interactants are well-known to have an impact on the pragmatic choices of contemporary language users, to date research has largely not considered these factors in earlier varieties of English. Recent exceptions to this are Elsweiler (2022) and Elsweiler and Ronan (2023). In these studies, the authors could show that social factors can be shown to have an influence (e.g. gender, Elsweiler 2022). Certain differences between varieties of English can be observed, yet, given not only restrictions in the amount of available data, but especially also of the available socio-demographic information, the creation of robust statistical evidence is an issue.

The present paper investigates address and leave taking formulae in approximately 200 letters written by Irish letter writers in the 18th century. The corpus is comprised partly of letters that are held in autographed letter collection at the National Library of Ireland (131 letters) and are available in unedited format on the webpages of the NLI. Further, 88 letters written by and to Irish emigrants in America were taken from the Corviz corpus. Sociodemographic data on letter writers and recipients is available in many, but not all cases. Where these are available, social status of letter writers and recipients are determined on information about occupation or titles, and, where known, the social relationship between the interactants is determined and coded. The data were then analysed with the help of regression analyses and decision trees (Weihs & Buschfeld 2021).

Results of the analysis indicate that pragmatic structures exhibit most variation and innovation in letters written within the same status group. Across status groups, both writing upwards and downwards, the pragmatic markers are both more formal and more stable.

References

"Autographed letters signed collection, ca. 1603–ca. 1972" held at the National Library of Ireland. https://catalogue.nli.ie, (last accessed 02 September 2023).

Corviz = Coriecor visualized. Eds. Amador-Moreno, Carolina P., Nancy E. Ávila-Ledesma, Karen Corrigan, María F. García-Bermejo Giner, Kevin McCafferty, Niall O'Leary, Javier Ruano García, Pilar Sánchez-García, Manuel Villamarín-González. Compiled at the University of Bergen. https://corviz.h.uib.no/index.php (last accessed 30 January 2023).

Elsweiler, Christine. 2022. Gender Variation in the Requestive Behaviour of Early Modern Scottish and English Letter-writers? A Study of Private Correspondence. *Journal of Historical Sociolinguistics* 8 (1), 55–88.

Elsweiler, Christine & Patricia Ronan. 2023. From *I am, with sincere regard, your most obedient servant* to *Yours sincerely*: The simplification of leavetaking formulae in 18th-century Scottish and Irish English letters. *ICAME Journal* 47 (1), 1–17.

Weihs, C. & Buschfeld, S. (2021). Combining Prediction and Interpretation in Decision Trees (PrInDT) – a Linguistic Example. ArXiv: http://arxiv.org/abs/2103.02336

Workshop 2: Diversity and Innovation in Concordance Organisation and Interpretation

Convenors: Stephanie Evert[1], Natalie Finlayson[2], Michaela Mahlberg[1] and Alexander Piperski[1]

([1]FAU Erlangen-Nürnberg, [2]University of Birmingham)

Corpus linguistics has come a long way since the first corpora were compiled for computer-assisted linguistic analysis of general language for lexicographic purposes. In recent decades, we have seen an abundance of technical innovations in quantitative approaches to managing the large volumes of data returned by searches in corpora of millions and billions of words, with automated analyses of frequency, collocation, and keyness providing invaluable overviews of patterns in language samples across registers, text types, and disciplines. Comparatively little work, however, has been done to enhance the qualitative methods and tools we use to explore the context around the words and phrases highlighted by quantitative procedures, or how the theories we draw upon to explain their patterns of use are integrated into the analytical process. The "bridge" between quantitative and qualitative corpus methods in corpus linguistics is the concordance, that is, the visualisation of the results of a corpus query in stacked lines of context to the left and right. Given the critical role of this vertical "reading" (Sinclair, 2003) of context in setting corpus linguistics apart from other computational approaches to language study (Hunston, 2022), the apparent neglect of innovation in concordancing is surprising – in terms of both methodology and software tools.

One of very few attempts to develop a systematic methodology for concordance interpretation is Sinclair's (2003) model of the concept of 'lexical item'. We might speculate that his approach, which usually focuses on collocational and colligational patterns and semantic preference/prosody, has shaped concordancing methods and the software available today. However, as concordance analysis becomes increasingly popular across linguistic and other text-based disciplines, a model grown from work in lexicography cannot account for the full range of purposes of a methodology characterised by a mix of qualitative and quantitative approaches set in a range of theoretical contexts.

The goal of our workshop is to start filling this gap by taking stock of qualitative concordance interpretation methodologies in different disciplines as well as current or proposed innovation in concordance tools. To this end, we invite a panel of international experts to discuss practices in their areas of specialism and reflect on how the hermeneutics of the concordancing process and use of tools may evolve in future. We present a workshop comprising papers and software demonstrations, covering topics such as concordance visualisation (Dietsch & Piperski), mathematical models of algorithms for organising concordances (Evert), tool-independent strategies for concordance reading (Finlayson & Mahlberg), applications of concordance reading (Hunston), updates to popular concordancing tools (Anthony), and others to be confirmed.

Concordancing in the twenty-first century: A brief review of current practices

Natalie Finlayson and Michaela Mahlberg (University of Birmingham, FAU Erlangen-Nürnberg)

At the beginning of the century, Teubert (2001: 125–126) warned of a downside to the rapid expansion of corpus linguistics, noting that an inward-looking focus on corpus construction and data standardisation may come at the expense of furthering "the original gain that the analysis of corpora may contribute to our knowledge of language." Not unrelatedly, Sinclair (2003) pointed to a need for reliable methodological procedures in anticipation of increasing amounts of concordancing work being carried out computationally. How such a framework might look in practice is currently unknown, but its development represents a crucial step towards moving the discipline forward in a time of renewed growth and technological change.

In this paper, we ask what still needs to be done to bring a level of systematicity to concordance reading that aligns with the flexibility and popularity of the approach and the technical innovations of the present day. As a starting point, we illustrate the variety of ways in which analysts select, organise, and interpret concordance data with examples from literature in four disciplines that bring different motivations and assumptions to the process: lexicography, data-driven learning, corpus-assisted discourse analysis, and literary stylistics. Our overview builds on a small body of work (e.g., Sinclair, 2003; Anthony, 2018; Gillings & Mautner, 2023; Hanks, 2013; Hoey, 2005; Hunston & Francis, 2000; Mahlberg, 2005) that lays the foundations for the development of a structured concordancing methodology based on principled choices about *what* information analysts want to see, *how* they want to see it, and *how* they will make sense of it. By mapping analysts' decisions and considering how their concordancing methods are driven by practical and theoretical contexts, our review not only enhances our understanding of trends that characterise disciplinary practices but also offers insights into three fundamental strategies that underpin concordancing work more broadly. Most strategies can be described as a means of creating a *subset* of data to be analysed, *ordering* concordances so that patterns can be revealed more easily, or *grouping* concordance lines in preparation for interpretation with reference to linguistic and other frameworks.

We envisage that discussions in today's workshop will build on these beginnings, paving the way for systematic, transparent, and much-needed theoretical, methodological, and technical innovation in each of the three areas identified.

References

Anthony, L. (2018). Visualization in corpus-based discourse studies. In C. Taylor and A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 197–224). Routledge.

Gillings, M. and Mautner, G. (2023). Concordancing for CADS: Practical challenges and theoretical implications. *International Journal of Corpus Linguistics* 29(1), 34–58.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. MIT Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.

Hunston, S. and Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins.

Mahlberg, M. (2005). *English general nouns: A corpus theoretical approach*. John Benjamins.

Sinclair, J. (2003). *Reading concordances*. Pearson.

Teubert, W. (2001). Corpus linguistics and lexicography. *International Journal of Corpus Linguistics* 6(1), 125–153.

Concordance analysis in CADS: Does "expanding the line" really work?

Mathew Gillings (Vienna University of Economics and Business)

Located at the intersection between quantitative and qualitative approaches to textual analysis, concordance analysis is one of the main techniques within a corpus linguist's toolkit. However, despite a growing body of work critically exploring previously unquestioned mainstays of corpus methods (Mautner, 2015; Taylor & Marchi, 2018; Gillings et al., 2023), it is rare to see this applied to concordance analysis specifically. One recent example of such work is Gillings & Mautner (2023), which explored the range of different issues that researchers may encounter when interpreting concordances within a corpus-assisted discourse analysis (CADS) framework. Drawing on an almost 20-million-word corpus of every article and book review published in *Administrative Science Quarterly* from 1956–2018 (Mautner & Learmonth, 2020), the paper identified eight key issues in concordance line interpretation: noise in the corpus, non-standard syntax, unclear referring expressions, unclear quotation source attribution, technical terms and jargon, acronyms and initialisms, unspecific co-text, and finally lines that are unrelated to the research question. Around one quarter of all lines analysed were uninterpretable; a number that is perhaps relieving or surprising, depending on what exactly one uses concordance analysis for.

For those who use concordance analysis to aid in (critical) discourse analyses specifically, this is likely to be surprising, and a problem. After all, the key remit is to get a sense of the range of different views and representations in a corpus, regardless of whether they are frequent or not. There are few solutions for what to do with uninterpretable concordance lines. Weisser (2016) suggests removing such lines from the analysis (provided such decisions are properly documented), whilst Collins (2019) suggests either extending the span of the co-text or revisiting the full text. These solutions are centred around either increasing the amount of co-text that is viewed or being openly transparent about removing them. Collins' advice to "expand the concordance line" is commonly cited in corpus linguistics literature.

This talk explores the extent to which this advice works in practice. Does "expanding the concordance line" really help? Returning to the uninterpretable lines identified by Gillings and Mautner (2023), I examine what additional steps are necessary to make them interpretable focussing on which of the eight key issues are potentially salvageable and which continue to be a problem. Preliminary analyses suggest that interpretability issues due to unclear referring expressions, unclear quotation source attribution, and unspecific co-text can often be solved by expanding the concordance line. Other lines, however, require further digging either elsewhere in the corpus, or from outside of it. The talk concludes with some thoughts on how developers of concordancing systems may aid (or indeed fix) these issues.

References

Collins, L. (2019). Corpus linguistics for online communication: A guide for research. Routledge.

Gillings, M. & Mautner, G. (2023). Concordancing for CADS: Practical challenges and theoretical implications. *International Journal of Corpus Linguistics* 29(1), 34–58 https://doi.org/10.1075/ijcl.21168.gil

Gillings, M., Mautner, G. & Baker, P. (2023). *Corpus-assisted discourse studies.* Cambridge University Press.

Mautner, G. (2015). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse studies* (3rd ed.) (pp. 154–179). SAGE.

Mautner, G., & Learmonth, M. (2020). From *administrator* to *CEO*: Exploring changing representations of hierarchy and prestige in a diachronic corpus of academic management writing. *Discourse and Communication* 14(3), 273–293.

Taylor, C., & Marchi, A. (2018). *Corpus approaches to discourse: A critical review.* Routledge.

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis.* Blackwell.

Modelling the output from concordance lines

Susan Hunston and Xin Susie Sui (University of Birmingham, Capital Normal University)

Although concordancing has a very long history, it is the Key Word in Context format of concordance lines that is associated with Corpus Linguistics. Concordance lines tend to be somewhat marginalised in Corpus Linguistics research, with their significance limited to (a) finding information of importance to lexicography, (b) checking the results of quantitative studies, and (c) finding examples of phenomena identified by other means. However, the output from studies of concordance lines has had a considerable impact on models of language that have either emerged from or been substantially influenced by the study of corpora.

The starting point for this paper is Sinclair's work in the 1980s that developed concepts such as the Unit of Meaning and the Idiom Principle (Sinclair, 1991: 2004). This work focused on lexis and grammar as a single system, on the unity of form and meaning, and on the location of meaning in the phrase rather than in the individual word. Sinclair demonstrated his approach in a series of specific word-studies (2003, 2004), and the Collins COBUILD series of dictionaries and grammars provided detailed descriptions of English using the same principles. The work was extended and given a further theoretical perspective, by, for example, Lexical Priming (Hoey, 2003), Local Grammar (Barnbrook, 2002; Cheng & Ching, 2016), and Corpus Pattern Analysis (Hanks, 2013). The scrutiny of concordance lines by individuals was the key methodology used in each case.

Sinclair, however, was not alone in recognising the interconnectedness of form and meaning, lexis and grammar. The concept of the Construction (Goldberg, 1995; Hoffman & Trousdale, 2013) developed independently of the Unit of Meaning, but is very similar to it, in particular in its rejection of the lexis-grammar distinction and its identification of meaning with form. Many of the examples of Units of Meaning discussed in the literature could be described as Constructions, and vice versa. The FrameNet project (Fillmore et al., 2003), with its mapping of meaning to form, shares much with the notion of Local Grammar, even though, again, they developed independently and largely unaware of each other. In consequence, there are multiple approaches that are similar but not identical, taking different theoretical standpoints and focusing on distinct but overlapping language phenomena. They all have a starting point in the scrutiny of large amounts of naturally-occurring language, with concordance lines at the heart of this.

This paper tries to make sense of this muddle of terminology and proposes an approach to thinking about four concepts – Units of Meaning, Local Grammar, FrameNet and Construction Grammar – that clarify what they share and how they differ. A series of oppositions is used to make these comparisons: mental focus vs output focus; form-to-meaning vs meaning-to-form; notion focus vs function focus; general vs partial theory; specific vs non-specific context. The result is a step-wise model that traces a progression of thinking from observation of concordance lines to contextualised theories of language.

Acknowledgement: This study is partially supported by the MOE Project of Humanities and Social Sciences (Project No. 19YJC740069) and the China Scholarship Council (File number: 202307300026).

References

Barnbrook G. (2002). *Defining language: A local grammar of definition sentences*. Benjamins.

Cheng, W., and Ching, T. (2016). 'Not a guarantee of future performance': The local grammar of disclaimers. *Applied Linguistics*, *39*(3), 263–301.

Fillmore C., Johnson C., and Petruck, M. (2003). Background to FrameNet. *International Journal of Lexicography* 16(3): 235–250.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations.* MIT Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.
Hoffman T. and Trousdale, G. (2013). *The Oxford handbook of construction grammar*. Oxford University Press.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
Sinclair, J. (2003). *Reading concordances*. Longman.
Sinclair, J. (2004). *Trust the text: Language, corpus, and discourse*. Routledge.

A sentence embedding approach to concordance searching and sorting

Laurence Anthony (Waseda University)

Concordancing has long been a cornerstone of corpus linguistics research, providing scholars with a powerful method to explore lexical and grammatical patterns in target corpora. It is also one of the most common approaches introduced to learners in a data-driven learning (DDL) classroom. Despite the strengths of the approach, it also suffers from two major limitations. Firstly, concordance searching requires the use of single or multi-word queries that are often fixed in nature and can quickly increase in complexity depending on the aim. For example, to account for possible variations in usage, these queries usually require the use of alternative options or the inclusion of in-word or between-word wildcards. If the researcher, teacher, or learner hopes to capture subtle variations in usage in the corpus (e.g., spelling differences between UK and US speakers, idiomatic expression with synonym variations, semantically equivalent words or phrases), these differences have to be recognized from the outset and accounted for in the query.

A second limitation of concordancing relates to the sorting of results. Typically, results are sorted alphabetically on the center (node) word, or words to the left or right of the node word. This ordering leads researchers, teachers, and learners to have to scan through all results to find relevant, salient patterns of usage. Recently, we have seen innovations such as KWIC patterns (Anthony, 2018, 2022) that calculate the frequency of occurrence of concordance result patterns and order the results accordingly. However, even here, if the query generates many thousands of hits for a particular pattern, there is still a need to sort these results in some meaningful way before they can be interpreted.

Over the past year, much attention has begun to focus on the potential impact of Artificial Intelligence (AI) on corpus research. In this paper, I introduce an innovative approach to concordance querying and sorting that integrates traditional concordance methods with transformer-based sentence (or sentence fragment) embeddings. Using sentence embeddings, I show how concordance search queries can be greatly simplified and also allow for more nuanced and context-aware analysis of linguistic phenomena than previously possible. In a case study using the BE06 (Baker, 2009) and AmE06 (Potts and Baker, 2012) corpora, I first demonstrate how traditional concordance queries can be interpreted in a “fuzzy” way, allowing subtle differences in language usage to be captured without the need for careful crafting of the query itself. Next, I show how an embedding model can be used to cluster the results of a traditional concordance analysis based on semantic similarity, leading to novel groupings and orderings of results. I then show how an embedding model can be used to match expressions in one language variety with those in another, leading to truly novel concordance analyses. The paper finishes with a discussion of future directions in AI and the potential impact on concordance tool development.

References
Anthony, L. (2018). Visualization in corpus-based discourse studies. In C. Taylor and A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 197–224). Routledge.
Anthony, L. (2022). What can corpus software do? In A. O'Keeffe and M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 237–276). Routledge
Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics* 14(3), 312–337.
Potts, A. and Baker. P. (2012). Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics* 17(3), 295–324.

A mathematical model of algorithms for organising concordances

Stephanie Evert (FAU Erlangen-Nürnberg)

An important step towards achieving transparency in concordancing is aligning the hermeneutics of the process with the computational algorithms available to support it. To connect algorithms and their combinations to the interpretative part of concordance reading, we propose a formal framework that systematises classes of algorithms based on their mathematical properties and determines how different algorithms can be combined. Our framework categorises algorithms into five strategies based on how they manipulate the concordance view displayed to the analyst:

(1) *Selecting* algorithms subset concordance lines, typically using metadata categories or manual selection (e.g. ranges of lines, or one or more of the sets formed by a partitioning or clustering algorithm, see below).
(2) *Sorting* algorithms rearrange concordances by comparing pairs of lines (A, B) to determine whether A should sort before B, B before A, or both are tied. A typical example would be to sort alphabetically by the right or left context of the node.
(3) *Ranking* algorithms also rearrange concordances, based on a numerical value assigned to each line, with the largest values shown at the top of the concordance view. Examples include readability scores, average word frequency, or number of salient collocates in the context.
(4) *Partitioning* algorithms divide concordances into sets of lines that share a certain observable feature. Such sets could consist of all lines from the same text genre or author, all lines where the token immediately to the left of the node has the same POS tag, or lines that have been manually categorised according to bespoke criteria. The criteria by which lines are partitioned also provides frequency counts for the property of interest (= sizes of the sets).
(5) *Clustering* algorithms collect concordance lines into hierarchically nested sets based on their mutual similarity (with a flat list of clusters as a special case). Examples include flat clustering based on lexical overlap or semantic similarity and a "POS tree" display that groups lines by the POS tag of the token to the right of the node at the highest level, then by the tag of the second token to the right, etc. Mathematically, clustering is represented by an ordered tree whose nodes correspond to sets of concordance lines.

Multiple sorting and ranking algorithms can be combined: the second algorithm breaks ties in the ordering of the first, the third breaks any remaining ties, etc. By contrast, only a single partitioning or clustering algorithm can be in effect because of potential conflicts between sets formed by different algorithms. This single partitioning or clustering algorithm determines the high-level organisation of the concordance, while lines within each set are ordered according to the sorting and ranking algorithms. Selecting plays a special role: it allows the analyst to "zoom in" on part of a concordance for more fine-grained analysis and forms a natural scope boundary. In this way, multiple partitioning and clustering algorithms can be used together in an analysis path, one after each selecting step.

Acknowledgement: This work has partially been funded by the *Deutsche Forschungsgemeinschaft* (*DFG*, German Research Foundation) – 508235423.

FlexiConc demonstrator: A front-end web app for structured concordance analysis

Levi Dietsch and Alexander Piperski (FAU Erlangen-Nürnberg)

FlexiConc is a software library developed to support a systematic approach to concordance analysis and interpretation by implementing existing and novel algorithms. It is not a comprehensive corpus management tool; rather, a 'concordance management tool' that can be integrated with other software. Testing and evaluating *FlexiConc* requires a front-end interface, which raises questions regarding the visualization of concordance reading strategies and the distribution of tasks between the front-end and back-end. In this talk, we will rationalize our design decisions and present a working version of the *FlexiConc* demonstrator.

The process begins when a user sends a query through the *FlexiConc* demonstrator to a host app, which could be any existing corpus management tool (e.g., *Corpus Workbench*, *CLiC*, *Sketch Engine*, *AntConc*). The host app returns the concordance data, which *FlexiConc* then passes to the library where users perform the required concordance operations.

Many corpus management tools (e.g., *CQPweb* and *Sketch Engine*) record procedural steps so that users can follow the sequence of operations performed and, if necessary, return to a previous step and continue from there. *FlexiConc* adopts a more intricate structure: the **operation-and-subset tree**, which facilitates complete research documentation. A set of concordance lines is represented as a node which can undergo various re-ordering (sorting and ranking), partitioning, and clustering operations. These are added as leaves attached to this node. Focusing on a subset of concordance lines—either through automatic selection or manual annotation—introduces a **scope boundary**. In terms of the tree, it is a node which can be further expanded with leaves by reapplying re-ordering, partitioning, and clustering operations. Nodes in the tree that are crucial for analysis can be marked as **snapshots** for later reproducible access by analysts or readers.

Figure 1 presents a prototype design for the *FlexiConc* demonstrator, including an operation-and-subset tree on the left. The current view (marked by an asterisk) selects lines from texts written in the 19th century and ranks them by the number of possessive pronouns in context.

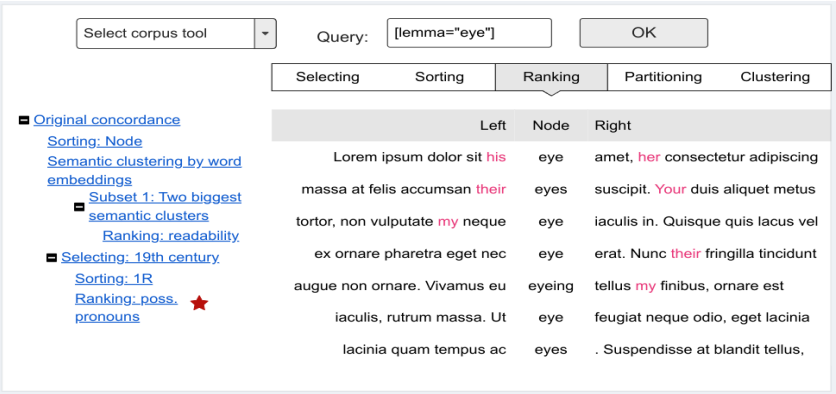


Figure 1. Prototype design for *FlexiConc* demonstrator.

The operation-and-subset tree effectively demonstrates how different concordance reading algorithms interact. When a user requests the application of an algorithm to a concordance view, two scenarios are possible:

- A child node is created from the current node (common when applying a Selecting algorithm).
- A sister node is formed, indicating either incompatibility with the current view or an override of the current algorithm. For example, Clustering algorithms are incompatible with each other; Ranking by readability, while technically compatible with Sorting by left context, adds new ordering scores with very few ties to the concordance lines, effectively overriding their previous order.

In summary, the purpose of the *FlexiConc* demonstrator is to illustrate a possible implementation of *FlexiConc* and present ways in which concordance analysis and interpretation can benefit from its features.

Acknowledgement: This work has partially been funded by the *Deutsche Forschungsgemeinschaft* (*DFG*, German Research Foundation) – 508235423.

Workshop 3: Interlocking Multilingual Corpora and Register(s): Diversity and Innovation

Convenors: Sylvi Rørvik and Marlén Izquierdo

(Inland Norway University of Applied Sciences, University of the Basque Country)

In accordance with long-standing ICAME tradition, we are pleased to welcome you to a pre-conference workshop focusing on contrastive linguistics on June 18, 2024 in Vigo, in connection with ICAME45. Contrastive workshops have been a recurrent feature at ICAME conferences for many years, and have been instrumental in furthering development in the field. The conference theme for ICAME45, "Interlocking Corpora and Register(s): Diversity and Innovation", provides an excellent opportunity to focus on diversity of registers in contrastive linguistics, and the workshop therefore has the related title "Interlocking Multilingual Corpora and Register(s): Diversity and Innovation".

We have particularly encouraged prospective participants to submit paper proposals that focus on multilingual investigations of lesser-explored registers, or that compare different registers or modalities. The ten papers included in the workshop all compare English with one or more of six other languages (Dutch, French, German, Norwegian, Swedish, and Spanish), and the authors investigate a wide range of linguistic features in a variety of registers, including fiction, promotional texts, blogs, conversation, academic prose, parliamentary discourse, and others.

References

Curry, N. & P. Pérez-Paredes. 2021. Stance nouns in COVID-19 related blog posts: A contrastive analysis of blog posts published in *The Conversation* in Spain and the UK. *International Journal of Corpus Linguistics* 26(4), 469-497.

Dupont, M. 2020. Placement patterns of English and French conjunctive adjuncts of contrast: The impact of register. *Languages in Contrast* 20(2), 263-287.

Ebeling, S.O. 2021. Minutes of action! A contrastive analysis of time expressions in English and Norwegian football match reports. In A. Čermáková, T. Egan, H. Hasselgård & S. Rørvik (eds), *Time in Languages, Languages in Time.* John Benjamins, 229-254.

Hasselgård, H. 2023. *Seem* and *appear* and their Norwegian verbal counterparts: A cross-register contrastive study. *English Studies* 104(1), 173-200.

Izquierdo, M. & M. Pérez Blanco. 2020. A multi-level contrastive analysis of promotional strategies in specialised discourse. *English for Specific Purposes* 58, 43-57.

Labrador, B. & N. Ramón. 2020. Building a second-language writing aid for specific purposes: Promotional cheese descriptions. *English for Specific Purposes* 60, 40-52.

Malá, M., D. Šebestová & J. Milička. 2021. The expression of time in English and Czech children's literature: A contrastive phraseological perspective. In A. Čermáková, T. Egan, H. Hasselgård & S. Rørvik (eds), *Time in Languages, Languages in Time.* John Benjamins, 283-304.

Werner, V. 2023. English and German pop song lyrics: Towards a contrastive textology. *Journal for language technology and computational linguistics JLCL* 36(1), 1-20.

Yuan, C. 2019. A battlefield or a lecture hall? A contrastive multimodal discourse analysis of courtroom trials. *Social Semiotics* 29, 645-669.

Selected list of publications from previous contrastive workshops

Aijmer, K. & H. Hasselgård (eds). 2015. *Cross-linguistic Studies at the Interface Between Lexis and Grammar*. Special issue of *Nordic Journal of English Studies,* (Vol 15:1). (ICAME34, Santiago de Compostela 2013)

Čermáková, A., S. O. Ebeling, M. Levin & J. Ström Herold (eds). 2021. *Crossing the Borders: Analysing Complex Contrastive Data*. *Bergen Language and Linguistics Studies* (BeLLS), Vol 11: 1. (ICAME41, Heidelberg 2020).

Čermáková, A., T. Egan, H. Hasselgård & S. Rørvik (eds). 2021. *Time in Languages, Languages in Time*. John Benjamins. (ICAME40, Neuchâtel 2019)

Čermáková, A., H. Hasselgård, M. Malá, & D. Šebestová (eds). Forthcoming/2024. *Contrastive Corpus Linguistics. Patterns in Lexicogrammar and Discourse*. Bloomsbury (ICAME42, Dortmund 2021)
Ebeling S.O. & H. Hasselgård (eds). 2015. *Cross-linguistic Perspectives on Verb Constructions*. Cambridge Scholars Publishing. (ICAME35, Nottingham 2014)
Ebeling S.O. & H. Hasselgård (eds). 2018. *Corpora et Comparatio Linguarum: Textual and Contextual Perspectives. Bergen Language and Linguistics Studies* (BeLLS) Vol 9: 1. (ICAME38, Prague 2017)
Egan, T. & H. Dirdal (eds). 2017. *Cross-linguistic Correspondences.* John Benjamins. (ICAME36, Trier 2015)
Janebová, M., E. Lapshinova-Koltunski & M. Martinková (eds). 2017. *Contrasting English and other Languages through Corpora*. Cambridge Scholars Publishing. (ICAME37, Hong Kong 2016)
Levin, M. & J. Ström Herold (eds). Forthcoming. Special issue of *Languages in Contrast*. (ICAME43, Cambridge 2022)

Please as a requestive marker: Contrasting its use and functions in English and Swedish comparable blog corpora

Karin Aijmer (University of Gothenburg)

The aim of the present study is to contribute to the study of the comparability of genres across languages by comparing the uses and functions of the requestive marker *please* in English and Swedish blog corpora. *Please* has been studied earlier in conversation (e.g. Wichmann 2004, 2005). Taking a cross-linguistic genre-based perspective means paying attention to how the characteristic features of the blogs can explain formal and functional properties of the expression across the compared languages.

Methodologically, the study is based on comparable corpora of English blogs (the Birmingham Blog Corpus (https://wse1.webcorp.org.uk/home/blogs.html) and Swedish blogs (included in the Korp Corpus https://spraakbanken.gu.se/korp/). Blogs are an emergent genre of internet communication sharing with informal conversation the feature of social interactiveness although the addressee (or addressees) is not physically present.

Two hundred occurrences of *please* with a function in the speech act of requesting are extracted from the two corpora and investigated in detail with respect to their syntax (position), type of sentence form and pragmatic function. *Please* in the Swedish blogs has been borrowed from English. A comparison will be made between *please* (in the Swedish blogs) and the distribution and use of the domestic *snälla* ('kind') in the blogs (cf. Ohlander 1985).

The preliminary findings based on a small number of examples indicate that *please* was used differently with regard to its position in the utterance and pragmatic function in the English and in the Swedish blogs. What characterizes the examples of *please* in the English blogs is that there is no indication that the blogger is personally involved or has a recipient for the blog in mind. *Please* was mostly placed in initial position (followed by an imperative) emphasizing the force of the request in a 'ritual' context where the blogger performs certain routine tasks (*please visit my website, please e-mail me*) in a situation where the rights and obligations of the blog user are clearly defined (cf. Kádár and House 2020). In the Swedish blogs *please* occurred most frequently in final position after many different syntactic types of requests (including single noun phrases) with the interactive function of making an appeal. The appeal may be to 'others', 'somebody', 'the weather gods' to give support, help or to comply with the blogger's request. The preliminary findings indicate that *please* and *snälla* are used in similar ways but differently than *please* in the English blogs.

References
Kádár, D.Z. and J. House. 2020 Ritual frames. A contrastive pragmatic approach. *Pragmatics* 30(1): 142-168.
Ohlander, S. 1985. 'Snälla ta med brickan!' Om ett nytt uttryck för hövlighet i svenskan. *Språkvård* 3: 4–15.
Wichmann, A. 2004. The intonation of *please*-requests: A corpus-based study. *Journal of Pragmatics* 36 (9): 1521-1549.
Wichmann, A. 2005. *Please* – from courtesy to appeal: The role of intonation in the expression of attitudinal meaning. *English Language and Linguistics* 9(2): 229-253.

What's in a title? A corpus-based contrastive analysis of titles in parascientific texts in English, French, and Spanish

Niall Curry (Manchester Metropolitan University)

Typically, a title is a reader's first introduction to a text. Across a range of texts and contexts, spanning newspaper headlines (Ifantidou, 2023), titles in novels (Martinez et al., 2016), and titles in academic research articles (Soler, 2009), research has shown that readers will often make the decision as to whether or not they will read a full text after reading its title. This readerly behaviour coupled with the metric-driven neoliberalisation of contemporary academia has greatly influenced how academics write, with more and more academics making use of so-called 'catchy' titles to engage readers and stand out amid the vast sea of research produced in our global publish or perish culture (Haggan, 2004). As such, in academic discourse research there has been a proliferation of studies centring on titles across a range of languages. These studies typically investigate the form and function of titles in well-established genres, such as research articles (e.g., Soler, 2009) with a view to better understanding how academics from different cultural backgrounds and disciplines engage their readers.

In academic discourse studies, for example, titles in soft science English language research articles have been found to make use of non-finite verbs and colons to create short and clear titles containing pre- and post-modifiers (Gómez et al., 1998). Conversely, titles in French academic research articles have been found to be ambiguous or unclear, impacting, for example, their categorisation in large international bibliographic databases (Alès et al., 2016), while Spanish titles in research articles appear to correspond to some degree with their English counterparts (Soler, 2009, 2011).

While titles in research articles have received much attention, the same cannot be said for titles in emerging parascientific genres, such as podcasts or blogs. This is somewhat surprising, as academics who produce blogs, for example, typically do so with the express purpose of disseminating research to a wider readership, often in different languages. Therefore, arguably, reader-engaging titles in blogs are of even greater importance when compared to research articles, which typically have a more captive audience. As such, from an academic discourse perspective, there is a need to better understand blog titles.

Notably, the genre of the academic blog remains somewhat unfixed and fuzzy (Curry & Pérez-Paredes, 2021), and there is an evident dearth of literature on academic blogs from cross-cultural, cross-linguistic, and cross-disciplinary perspectives. To shed light on how titles are used to engage readers in blogs written by academics and scientists, this paper presents a corpus-based contrastive analysis of titles in academic blogs in English, French, and Spanish. Using a corpus of academic blogs from the Conversation UK, France, and Spain, the study offers a taxonomical description of title forms and functions in blogs in English, French, and Spanish.

Overall, the findings offer insights into titular variation across discipline and language, highlighting cultural variation in how writers engage readers through titles in parascientific texts.

References

Alès, C., Arena, R., Brandt-Grau, A., Chaabane, N., Cortes, G., Crespin, R., Didier, E., Fretel, J., Gardey, D., Guermeur, I. and Gueye, L. (2016) *Rapport de prospective du Conseil scientifique de l'Institut des sciences humaines et sociales du CNRS*. Paris: Centre National de la Recherche Scientifique.

Curry, N. & Pérez-Paredes, P. (2021) Stance nouns in COVID-19 related blog posts: A contrastive analysis of blog posts published in The Conversation in Spain and the UK. *International Journal of Corpus Linguistics* 26(4), 469-497. https://doi.org/10.1075/ijcl.21080.cur

Gómez, I.F., Gómez, S.P., Silveira, J.C.P. and Garcia, J.F.C., (1998) Disciplinary variations in the writing of research articles in English, in Fortanet, I. and Dudley-Evans, T., eds., *Genre studies in English for academic purposes*. Valencia: Universitat Jaume I Press, 59–78.

Haggan, M. (2004) Research paper titles in literature, linguistics and science: Dimensions of attraction. *Journal of Pragmatics* 36, 293–331.

Ifantidou, E. (2023) Newspaper headlines, relevance and emotive effects. *Journal of Pragmatics* 218, 17-30.

Lettrilliart, L. and Schott, A.M. (2007) Rédiger et publier un article de recherche en médecine générale. *La Revue du Praticien* 21, 629–632.

Martinez, M., Stier, C., & Falcon, L. (2016) Judging a book by its cover: An investigation of peritextual features in Caldecott award books. *Children's Literature in Education* 47, 225-241.

Nagano, R. (2015) Research article titles and disciplinary conventions: A corpus study of eight disciplines. *Journal of Academic Writing* 5(1), 133–144.

Soler, V. (2009) Títulos científicos en lengua española: Estudio exploratorio. *Lebende Sprachen* 54(2), 50–58.

Soler, V. (2011) Comparative and contrastive observations on scientific titles written in English and Spanish. *English for Specific Purposes* 30(2), 124–137.

A cross-linguistic cross-register study of the verb phrase in English vs. Norwegian face-to-face conversation and fictional dialogue

Signe Oksefjell Ebeling (University of Oslo)

This paper investigates the verb phrase in English and Norwegian real (face-to-face) dialogue and fictional dialogue with the aim of establishing to what extent fictional dialogue "resembles real-life conversation" (Nykänen & Koivisto 2016: 3) in the two languages. The study is inspired by previous cross-register contrastive studies between English and Norwegian (e.g. Ebeling & Ebeling 2020; Ebeling Forthc.), as well as studies on the language of fiction vs. conversation (e.g. Biber et al. 1999; Leech & Short 2007; Jucker 2021). Such studies have uncovered differences (and similarities) both within and across languages and registers; however, the two registers of interest here have yet to be examined in an English-Norwegian contrastive perspective.

In a study of "features of orality" in the language of fiction, e.g. discourse and hesitation markers, Jucker (2021) found that such features, typical of face-to-face interaction, were more prominent in performed fiction (soap operas, movies and TV shows) than in written fiction. These findings inspired the current investigation of a more general linguistic feature in written fiction compared to face-to-face conversation, namely the verb phrase. Thus, the study focuses on a linguistic unit not considered a typical feature of orality. To make the comparison to conversation as fair as possible, only dialogic passages in written fiction will be consulted.

Against this backdrop, the study addresses the following research questions:

- To what extent are verb phrases in English and Norwegian formally similar in the two types of dialogue?
- To what extent do the two types of dialogue opt for semantically similar main verbs in the two languages?

The material is drawn from two sources: the English-Norwegian Parallel Corpus (ENPC) for fictional dialogue and from the International Comparable Corpus (ICC) for face-to-face conversations. To get a homogeneous and balanced dataset, ten VPs from each of the 20 (original) general fiction texts in the ENPC are extracted and analysed, along with ten VPs from each of 20 randomly selected conversations from the English and Norwegian components of the ICC. The main verbs in the 800 VPs are registered and classified semantically and the full VPs are classified according to their formal complexity, involving features such as tense and finiteness, aspect, voice and modality. Example (1) from English fiction shows a complex VP with the main verb TELL in the present perfect, while example (2) from Norwegian conversation shows a simple VP with the main verb VÆRE 'be' in the present tense.

(1) "I'*ve told* Jill to lie down," (ENPC-EN – fiction: AB1)
(2) det er jo vi vi *er* jo en ganske forskjellig type folk da (ICC-NO – conversation: S1A-021)
'it is of course we we *are* of course quite different types of people you know'

Preliminary observations suggest that both English and Norwegian mainly use relatively simple verb phrases, and slightly more so in conversation than in fiction, while semantically richer verbs tend to be used more frequently in fiction than in conversation in both languages. Norwegian conversation, in particular, seems to rely heavily on the semantically more basic lexical verb VÆRE 'be' (in 67 out of the 200 instances). Thus, this initial, and far from complete, analysis of the data suggests that even a general and essential linguistic element like the VP may serve as a distinguishing feature between fiction dialogue and conversation, and that writers only to a limited degree seem to aim (or manage) to imitate the verb phrase behaviour of spontaneous conversation in their fictional dialogues. To some degree these tentative results also corroborate previous cross-linguistic, cross-register findings in that register seems to be a more decisive factor than language (alone) regarding lexico-grammatical behaviour, at least between two closely related languages such as English and Norwegian.

References

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

Ebeling, Signe Oksefjell. Forthc. Structural and semantic features of adjectives across languages and registers. To appear in *Languages in Contrast* 2024.

Ebeling, Signe Oksefjell and Jarle Ebeling. 2020. Dialogue vs. narrative in fiction: A cross-linguistic comparison. *Languages in Contrast* 20(2), 288–313.

Jucker, Andreas. H. 2021. Features of orality in the language of fiction: A corpus-based investigation. *Language and Literature* 30(4), 341-360.

Leech, Geoffrey and Mick Short. 2007 [2nd ed.]. *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. Harlow: Pearson Longman.

Nykänen, Elise and Aino Koivisto. 2016. Introduction: Approaches to Fictional Dialogue, *International Journal of Literary Linguistics*, Vol. 5(2), Art. 1.

A corpus-based contrastive and translation study of EN *absolutely*, FR *absolument* and DU *absoluut*

Lobke Ghesquière, Gudrun Vanderbauwhede and Simon Copet (University of Mons)

Our paper presents a synchronic contrastive and translation study of the English adverb *absolutely* and its French and Dutch counterparts *absolument* and *absoluut*, respectively. Whereas EN *absolutely* has already received considerable attention, both from a monolingual (e.g. Aijmer 2016, Núñez Pertejo 2013, Partington 2004, Tao 2007) and a contrastive perspective (e.g. Aijmer 2020 on English and Swedish, Bardas 2008 on English and Norwegian, Carretero 2010 on English and Spanish), the French and Dutch adverbs have been less popular an object of study. Exceptions are Klein (1998), who analysed *absoluut* and other degree adverbs in Dutch, and Molina (2014), who discussed the French adverb *absolument* in his study on the negation of adverbs in *-ment*.

The aim of our study is twofold. Analysis of monolingual corpus data will allow fine-grained description of the pragmatico-semantic and syntactic features of the different uses of these adverbs in the three languages, English, Dutch and French, while analysis of parallel corpus data will enable identification of the strategies used by translators to render these adverbs in the other languages. As such this study aims to build on and extend the existing literature on these adverbs by widening the scope to less researched languages and by including translation data.

Monolingual data are drawn from the spoken component of the BNC2014 corpus, the Orféo corpus and OpenSoNaR. Parallel data are extracted from the sentence-aligned English, French and Dutch subcorpora of the Europarl-direct corpus (Cartoni et al. 2013), using SketchEngine (Kilgariff et al. 2014). We have chosen to use spoken data (or written-to-be-spoken for the parallel data) as this allowed us to find both (sub)modifier and independent uses of the adverbs. Written corpora are less likely to, for instance, contain occurrences of these adverbs as answers to questions (Tao 2007).

The qualitative analysis of the data is based on pragmatic, semantic and syntactic parameters. At the pragmatic level, register and context are analysed, which according to Tao (2007) and Núñez Pertejo (2013) influence the use of *absolutely* when it is used alone. At the semantic level, for instance, the adjectives modified by *absolutely*, *absolument* and *absoluut* are categorized following Lorenz (1999) and Bardas (2008). The polarity of the adjectives is also taken into account, as it could influence the translation choices made. At the syntactic level, following Tao (2007) and Carretero (2010), the nature of the word modified by the adverb is also tagged for. Cross-referencing of the data from the parallel corpora allows us to identify translation trends and confirm or refute our hypotheses.

Preliminary results seem to indicate that English *absolutely* behaves quite differently from its Dutch and French counterparts. Unlike *absoluut* and *absolument*, *absolutely* is not found in the data sets as a modifier of a negation markers (*not/no*). Moreover, whereas the French and Dutch adverbs are found to modify both epistemic and deontic modal auxiliaries, *absolutely* is not. In the English data sets, we also observe far less modification of elements conveying a modal meaning (e.g. *necessary*) and no instances even were found of *absolutely* modifying an element conveying volition. Finally, in terms of the polarity of the modified element, *absolutely* seems to have no clear preference for either positive or negative meanings, whereas *absolutement* and especially *absoluut* have a clear preference for negative collocates.

References

Aijmer, K. 2016. 'You're absolutely welcome, thanks for the ear': The use of *absolutely* in American soap operas. *Nordic Journal of English Studies* 15(2), 78.

Aijmer, K. 2020. Contrastive pragmatics and corpora. *Contrastive Pragmatics* 1(1), 28-57.

Bardas, A. C. 2008. *Amplifiers in English and Norwegian:* Absolutely, completely, entirely, perfectly, *and* totally *and their Norwegian correspondences: A study based on the English-Norwegian parallel corpus* [Master thesis].
Carretero, M. 2010. *You're absolutely right!* A corpus-based contrastive analysis of *absolutely* in British English and *absolutamente* in Peninsular Spanish, with special emphasis on the relationship between degree and certainty. *Languages in Contrast* 10(2), 194-222.
Cartoni B., Zufferey, S. & Meyer, T. 2013. Using the Europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics* 27, 23-42.
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. 2014. The Sketch Engine: Ten years on. https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf, (03/09/19)
Klein, H. 1998. *Adverbs of degree in Dutch and related languages*. Benjamins.
Lorenz, G. 1999. *Adjective intensification – learners versus native speakers. A corpus study of argumentative writing*. Rodopi.
Molina, J. V. 2014. De la négation de certains adverbes en *-ment*. *Cahiers de praxématique* 62.
Núñez Pertejo, P. 2013. From degree adverb to response token: *absolutely* in Late Modern and Contemporary British and American English. *Neuphilologische Mitteilungen* 114(2), 207-235.
Tao, H. 2007. A corpus-based investigation of *absolutely* and related phenomena in spoken American English. *Journal of English Linguistics* 35(1), 5-29.

Temporal catenatives in English and Norwegian (fiction, "non-fiction" and academic prose)

Hilde Hasselgård (University of Oslo)

The proposed study concerns what may be called temporal catenatives, such as *begin* and *continue* complemented by a verb. The most important selection criteria are that the two verbs should (i) share their subject, and (ii) be perceived as denoting a single action, as in (1) and (2); see Halliday & Matthiessen (2014: 567).

(1) It had *started to rain*. (BOE1)
Det hadde *begynt å regne*. (BOE1T)
(2) Jeg har *sluttet å spørre* for lenge siden. (EFH1)
I *stopped asking* a long time ago. (EFH1T)

The cross-linguistic comparison concerns English and Norwegian, as manifest in three registers: fiction, "non-fiction" and academic prose. The material comes from the English-Norwegian Parallel Corpus (ENPC) and the KIAP corpus (Cultural Identity in Academic Prose). The non-fiction part of the ENPC contains a mix of registers whereas the register of the KIAP subcorpus used is more homogenous, namely published research articles within economics. The ENPC is used primarily as a comparable corpus of original texts, but translations will be used to illuminate cross-linguistic differences in the use of catenatives as a component of the verb phrase.

The following questions are addressed:

- Do English and Norwegian use similar temporal catenatives, and with similar frequencies?
- Are temporal catenatives used differently across the three registers under study?

The catenatives were identified by searching for patterns in both languages: Verb + *to* and Verb + *-ing* participle in English, and Verb + infinitive in Norwegian, which uses only this verb form after catenatives (Holmes & Enger 2018). The resulting set of temporal catenatives denote the

beginning, continuation and end of activities (temporal phase, according to Halliday & Matthiessen 2014; aspect constructions in Egan 2008), e.g. English *begin, start, continue, keep, stop, cease* and Norwegian *begynne, starte, fortsette, slutte, stoppe, holde på (med).*

As Norwegian lacks a grammaticalized progressive aspect, an initial hypothesis was that Norwegian might favour temporal catenatives to mark continuation. This turned out not to be the case, however. Both languages use temporal continuatives mostly for inception/inchoation, particularly the cognates *begin* and *begynne*. Catenatives marking the end of an activity are least frequent in both languages, and less frequent in Norwegian than in English. Generally, the registers seem to differ more than the languages. As hypothesized, temporal catenatives are more common in fiction than in academic prose, and the mixed-register ENPC non-fiction is somewhere in between. Further analysis will also consider textual variation in order to find out how consistent the use of catenatives is across the registers.

References

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

The KIAP corpus (Cultural Identify in Academic Prose): www.uib.no/fremmedsprak/23107/kiap-korpuset

Egan, Thomas. 2008. *Non-finite Complementation. A Usage-based Study of Infinitive and* -ing *Clauses in English.* Rodopi.

Halliday, M.A.K. & Matthiessen, Christian M.I.M. 2014. *Halliday's Introduction to Functional Grammar*. Routledge.

Holmes, Philip & Enger, Hans-Olav. 2018. *Norwegian. A Comprehensive Grammar*. Routledge.

English-Spanish promotional texts and essays: Verbal use as a register marker?

Rosa Rabadán and Noelia Ramón (University of León)

Variation across registers has been investigated in detail, focusing on different aspects of discourse (Biber 1995; Zhang 2016; among others). Most of these studies have dealt with differences between spoken and written registers (Biber 1988) or have paid attention to particular registers, including academic writing (Hyland 1998, 2005). However, register differences across languages have not received the same attention. This paper investigates verbal use in two registers in English and Spanish: promotional discourse in the field of food and drink (specialized language) and essays (non-specialized language). The contrast is both intra- and cross-linguistic and has two aims: a) identify verb use similarities and differences between registers in the same language (Biber and Zhang 2016, 2018; Biber and Egbert 2018; Biber and Seoane 2021; Pontrandolfo 2021; Biber and Egbert 2023; Calvi et al. 2023) and across languages (Rabadán 2006, 2009, 2023), and b) check whether the findings qualify as register markers.

Our data come from three corpora: English (772,953 w) – Spanish (776,100 w) comparable CLANES, which contains original promotional texts in the food and drink industry (2015-2023) https://actres.unileon.es/wp/es/corpus-comparables/. For the non-specialized register, we use the English essays subcorpus in P-ACTRES 2.0 (514,786 w) https://actres.unileon.es/wp/es/corpus-paralelos/, and a Spanish corpus of essays built from CORPES XXI resources (492, 244 w) https://www.rae.es/corpes/. Both corpora feature contemporary materials. In the case of Corpes XXI, three field areas have been chosen to make it as similar as possible to the Essay contents in P-ACTRES 2.0: Current topics, Social Sciences, and Science and technology.

We focused on past and present tenses, imperatives, and modal verbs/ periphrases to obtain empirical data for this study. We used high-frequency lexical verbs for each of the corpora. In the specialized comparable corpus we queried *make, use, add, heat* and *serve*/ *hacer, añadir, tener, dejar,* and *poner*. In the non-specialized subcorpora, we queried *make, see, get, take,* and *go*/ *tener, ir, hacer, decir*, and *ver* for tenses and the imperative mood. English modal verbs and the conjugated Spanish modal periphrases (*poder, deber + inf*) were processed separately. The chi-square statistic was used to compute statistical significance at $p < .05$ in the two registers of the same language and cross-linguistically.

Preliminary results show that, in English, present tenses and imperatives are significantly more common in promotional discourse. By contrast, past tenses occur more often in essays than specialized texts. In Spanish, our data show that imperatives are also significantly more common in promotional discourse; present and past tenses occur more frequently in the non-specialized register.

Cross-linguistic preliminary results show that only the past tense does not present statistically significant differences in the specialized corpus. This means that past forms do not qualify here as register markers. The difference in the use of modal auxiliaries/periphrasis is statistically significant between the two registers in both Spanish and English, being more frequent in the non-specialized register.

These tentative results suggest that some tense and mood choices may be suitable markers to characterize registers grammatically. The procedure can be replicated with other registers to characterize them grammatically.

References

Biber, D. 1988. *Variations across speech and writing.* Cambridge University Press.

Biber, D. 1995. *Dimensions of register variation: A Cross-linguistic Comparison.* Cambridge University Press.

Biber, D. & M. Zhang. 2016. Register variation on the searchable web: a multidimensional analysis. *Journal of English Linguistics* 44: 95-137.

Biber, D. & M. Zhang. 2018. Expressing evaluation without grammatical stance: Informational persuasion on the web. *Corpora* 13(1): 97-123.

Biber, D., & J. Egbert. 2018. *Register variationoOnline*. Cambridge: Cambridge University Press.

Biber, D., & J. Egbert. 2023. What is a register? Accounting for linguistic and situational variation within – and outside– textual varieties. *Register Studies* 5(1): 1–22

Biber, D., & E. Seoane (Eds.). 2021. *Corpus-based approaches to register variation*. Amsterdam: John Benjamins. https://doi-org.unileon.idm.oclc.org/10.1075/scl.103

Calvi, M.V., G. Mapelli, M.C. Bordonaba & J. Santos. 2023. *Las lenguas de especialidad en español.* Roma: Carocci. https://hdl.handle.net/2434/1022108

Hyland K. 1998. Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics* 30: 437-455.

Hyland, K. 2005. *Metadiscourse: Exploring interaction in writing.* Blooomsbury.

Pontrandolfo, G. 2021. *Lingüística textual y discursos de especialidad: perspectivas de análisis*. Madrid: Arco Libros.

Rabadán, R. 2006. Modality and modal verbs in contrast: Mapping out a translation (ally) relevant approach English-Spanish. *Languages in contrast* 6(2): 261-306.

Rabadán, R. 2009. The present tenses in English and in Spanish: A corpus-based approach to cross-linguistic meaning and 'grammatical transfer.' In *Cognitive Approaches to Language and Linguistic Data. Studies in Honor of Barbara Lewandowska-Tomaszczyk*. Wiesław Oleksy, P. Stalmaszczyk (eds). Peter Lang.

Rabadán, R. 2023. Light verb constructions in English-Spanish translation. In *Corpus use in cross-linguistic research: Paving the way for teaching, translation and professional communication.* M. Izquierdo and Z. Sanz-Villar (eds). John Benjamins, pp. 34-50.

Zhang, M. 2016. A multidimensional analysis of metadiscourse markers across written registers. *Discourse Studies* 18(2): 204–222. http://www.jstor.org/stable/24815288

Extended premodifiers in English and German fiction and non-fiction

Jenny Ström Herold and Magnus Levin (Linnaeus University)

This contrastive, corpus-based register study explores the frequencies and forms of extended premodifiers in English and German fiction and non-fiction. We here define extended premodifiers (German 'erweiterte Attribute', cf. Magnusson 1995: 172) as consisting of an adjectival or participial head, having one or many 'extensions', which may function as arguments or adjuncts:

(1) a *relatively shady* area (LEGS, non-fiction)
(2) ein *mit Backpapier ausgelegtes* Blech ['a with wax-paper lined baking-pan'] (LEGS, non-fiction)

Dean (1971:230) suggests that "the extended premodifier, at least in the traditional sense of the term, is definitely German, not English". For German, Fagan (2009:125) gives the example *ein in der amerikanischen und europäischen Wirtschaft inzwischen weit verbreitetes Instrument* ['an in the American and European economy meanwhile widely spread instrument'], showcasing their potentially high complexity. Nevertheless, extended premodifiers are occasionally – albeit briefly – addressed in English grammars as well. For instance, according to Huddleston & Pullum (2005: 119), adverb extensions seem to be common in English (*extremely hot day*), but other categories may also occur – prepositional phrases (*an in some respects good idea*) and noun phrases (*two hours long trip*). While the German extended premodifier is fairly well described and researched (e.g., Solfjeld 2004; *Duden, die Grammatik* 2009: 563−566; Ström Herold & Henriksson 2022), this is not the case for English.

Contrastive German-English studies on this construction type are almost non-existent, a rare exception being Fabricius-Hansen (2010). Her mostly introspective study indicates that German has a greater tolerance for extended premodifiers than English, English mainly being limited to adverb extensions (see example (1)). On a more general level, studies have shown that German is less inclined to using postmodification than English (e.g., Teich 2003: 183) but also that premodifiers are more common in "expository registers" (i.e., newspapers and academic) – rather than fiction and conversation – in English (Biber et al. 2021 [1999]: 591).

The data for our study come from two different corpus collections. The non-fiction material was collected from the Linnaeus University English-German-Swedish corpus (LEGS), and the fiction material comes from the English-Swedish Parallel Corpus (ESPC) and the Oslo Multi-lingual Corpus (OMC). The LEGS data consists of, e.g., popular science and self-help books from the 2010s, while the sampled parts of the ESPC and OMC comprise English and German original fiction from the 1980s and 1990s.

Based on the above-mentioned studies, it is reasonable to assume that these constructions are more common, more complex and varied in German and in the non-fiction register than in English and fiction. These hypotheses gain some preliminary support from our pilot data: the rank order of frequencies for the registers follows the expected pattern, with extended premodifiers being the most frequent in German non-fiction and the least in English fiction. Also, the English instances identified are of the "minimal" kind with adverb extensions in both fiction (*a very low rent*) and non-fiction (*newly acquired resources*). Less evident, however, are at present the differences between the German registers, both as regards frequency and complexity. The frequency does not differ very much, and more complex instances, such as (3) below, which contains two extensions – a prepositional phrase (*mit dem Messer*) and an adverb phrase (*scharf*) – are not much rarer in fiction than in non-fiction:

(3) der wie *mit dem Messer scharf geschnittenen* Haaransatz ['the as with the knife sharply cut hairline'] (OMC, fiction)

It remains to be seen whether this is due to the relatively small material sampled so far, or perhaps a difference in time periods for the registers, the OMC fiction data being about 30 years older than the LEGS non-fiction data.

References

Biber, Douglas, Johansson, Stig, Leech, Geoffrey Conrad, Susan & Edward Finegan. 2021 [1999]. *Longman grammar of spoken and written English*. Harlow: Longman.

Dean, O. C., Jr. 1971. The extended modifier: German, not English. *American Speech* 46 (3–4): 223–230.

Duden – die Grammatik (Volume 4). 2009. Mannheim: Dudenverlag.

Fabricius-Hansen, Cathrine. 2010. Adjektiv-/Partizipialattribute im diskursbezogenen Kontrast (Deutsch–Englisch/Norwegisch). *Deutsche Sprache* 38(2): 175–197.

Fagan, Sarah M. B. 2009. *German: A linguistic introduction*. Cambridge: Cambridge University Press.

Huddleston, Rodney & Pullum, Geoffrey K. 2005. *A student's introduction to English grammar.* Cambridge: Cambridge University Press.

Magnusson, Gunnar. 1995. Deutsch–Schwedisch kontrastiv. Stolpersteine bei avancierter Übersetzung. *Moderna språk* 89(2): 164–179.

Solfjeld, Kåre. 2004. Zur Wiedergabe deutscher erweiterter Attribute in authentischen norwegischen Übersetzungen. *HERMES – Journal of Language and Communication in Business* 33: 89–115.

Ström Herold, Jenny & Henriksson, Henrik. 2022. Angekommen im Schwedischen? Deutsche Partizipialkonstruktionen in schwedischer Übersetzung. *Moderna språk* 116(1): 67–97.

Teich, Elke. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin: De Gruyter Mouton.

Mirativity in exclamative constructions: A cross-linguistic and cross-register approach

Faye Troughton (University of Mons)

This study provides a comparison of mirativity in English and French exclamative constructions in parliamentary discourse and written fiction. Exclamative constructions in English and French are generally accepted as including instances such as (1) to (6). Exclamative constructions are fronted by interrogative words and, in matrix position, are distinguishable from interrogatives by the absence of subject-auxiliary inversion. These constructions are understood to appear in independent, verbless, and embedded realizations. Pragmatically, they are characterized by their conveying of presupposed content, subjectivity, high degree, and mirativity (denoting surprise or exceeded expectations) (cf. Michaelis & Lambrecht 1996; Delancey 1997; Michaelis 2001; Rett 2008, 2011; Krawczak & Glynn 2015; Unger 2019). It is this final characteristic that concerns this study.

(1) *Oh, my word, what a sight she is!* (WB Brbooks)

(2) *Oh God, how terrible.* (WB Brbooks)

(3) *You've no idea how good it is to see a friendly face.* (WB Brbooks)

(4) *Mais quelle idiote je suis, pense -t-elle à présent.*
But what idiot I am thinks.0.she at present.
'But what an idiot I am, she now thinks.'
(PORTANTE Jean, Mrs Haroy ou la mémoire de la baleine, 1993)

(5) *Que de drames humains!*
That of drama human
'So many human tragedies!'
(Europarl-direct, Speaker ID 66, Martin, Hugues, PPE-DE)

(6) *Je remarquai combien son sourire un peu cruel*
I remarked how.much their smile a little cruel
était séduisant.
was seductive.
'I noticed how seductive his slightly cruel smile was.'
(GRACQ Julien, Le Rivage des Syrtes, 1951)

Krawczak & Glynn's (2015: 354) operationalization considers mirativity the simultaneous instantiation of "conceptual incongruity" and "functional performativity", and a scalar phenomenon, which is the line taken here. Conceptual incongruity concerns the degree to which an event or situation is incongruent or surprising and considers the immediacy of the surprise along with other contextual factors (Krawczak & Glynn 2015: 361). Functional performativity is the "enactment of the state of incongruity", or how this is portrayed through elements of language. This may include capitalization, punctuation, and more generally other elements of language that make an utterance more emotive: repetition, interjections, and elements that heighten "addressee-orientation" (Krawczak & Glynn 2015: 363). It is considered here that, specifically in the case of exclamatives, the realization of the construction may come into play as well. Neveux (2018: 205) argues that "a completed exclamative structure loses in expressivity what it gains in analysis, that the essence of exclamation rests in the beginning of the structure, in the *Wh*-phrase". If a sense of "surprise" is part of this expressivity, this would imply mirativity is stronger in verbless exclamative constructions, less so in the independent yet full exclamatives, and weak or absent in embedded exclamatives.

This study compares the conceptual incongruity, functional performativity, and clause variation (independent, embedded, or verbless) across English *what* and *how* and French *quel*, *combien (de), que (de),* and *comme* exclamatives in two registers: written fiction and parliamentary discourse. For the former, samples of 100 were taken from both the Wordbanks British books (HarperCollins 2009) and Frantext RL-1950+ (ATILF) subcorpora, and for the latter, exhaustive extractions were made from the Europarl-direct directional subcorpora (Cartoni & Meyer 2012). While all exclamatives show low mirativity generally, French exclamative constructions indicate higher mirativity across both registers.

References

DeLancey, Scott. 1997. Mirativity: The grammatical marking of unexpected information. *Linguistic Typology* 1(1), 33-52.

Cartoni, Bruno and Thomas Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odjik & Stelios Piperidis (eds.) *Proceedings of 8th International Conference on Language Resources and Evaluation (LREC),* Istanbul: European Language resources Association, 2132-2137.

Frantext, ATILF, Nancy, 1998-2023, https://www.frantext.fr (last access 20/12/2023).

HarperCollins. 2009. WordBanksOnline. https://wordbanks.harpercollins.co.uk/release_notes/Word-BanksOnline_English.html (last access 20/12/2023).

Krawczak, Karolina & Dylan Glynn. 2015. Operationalizing mirativity: A usage-based quantitative study of constructional construal in English. *English Review of Cognitive Linguistics* 13(2), 353-382.

Michaelis, Laura A. & Knud Lambrecht. 1996. Toward a construction-based theory of language function: The case of nominal extraposition. *Language* 72(2), 215-247.

Michaelis, Laura A. 2001. Exclamative Constructions. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.) *Language typology and language universals* 2. Berlin: De Gruyter, 1038-1050.

Neveux, Julie. 2018. Grammar and feelings: a study of *wh*-exclamatives in Katherine Mansfield's short stories. *Etudes de stylistique anglaise* 12, 193-222.
Rett, Jessica. 2008. A degree account of exclamatives. In Tova Friedman & Satoshi Ito (eds.) *Proceedings of Semantics and Linguistics Theory XVIII.* Ithaca/New York: CLC Publications.
Rett, Jessica. 2011. Exclamatives, degrees and speech acts. *Linguistics and Philosoph.* 34, 411-442.
Unger, Christoph. 2019. Exclamatives, exclamations, miratives and speaker's meaning. *International Review of Pragmatics* 11, 272-300.

Investigating the role of *not*-fragments in colloquialisation

Laura Abalo-Dieste (University of Vigo)

This study investigates the relation between *not*-fragments and colloquialisation in English. Colloquialisation, coined by Mair (1997), refers to the increased preference for informal linguistic strategies over time in writing (e.g. phrasal verbs and contractions). Fragments have received scholarly attention (see, for example, Bowie & Aarts 2016) as syntactically incomplete or non-canonical structures that seem to contribute to interactional discourse as full propositional sentences (e.g. *if only it would!* and *Hi to Simon*). Even though fragments are not uncommon in written language (see, for example, Fernández-Pena 2021), they are "far more common in conversation than in the written registers" (Biber et al. 1999: 225). A bias towards informal English supports the idea that colloquialisation can be observed through the frequency and distribution of fragments across registers at different levels of (in)formality. This paper widens the spectrum of colloquialisation strategies by assessing whether *not*-fragments, in particular, signal colloquialisation in contemporary English.

Not on my watch and similar structures (e.g. *no way*, *not likely* and *not in a million years*) have been identified as "idiomatic negative answers" (Huddleston & Pullum 2002: 849) or "*not*- fragments" (Capelle 2020). Negation with *not* – unlike the more formal variant with *no*-negation – in complete sentences has been found to be pervasive in spoken language and linked to the process of colloquialisation in British English (Smitterberg 2021: §5). This paper investigates the recent diachrony of English on the premise that colloquialisation, whereby "informal options which have been available for a long time are chosen more frequently today than would have been the case thirty years ago" (Mair 1997: 203), may also explain the use of *not*-fragments. More specifically, it explores the occurrence and evolution of nominal *not*-fragments in both spoken and written English with the purpose of demonstrating whether *not*-fragments exhibit a bias towards informal speech contexts and an upsurge in frequency, thus indicating a link to colloquialisation. To that end, the frequency and textual distribution of *not*-fragments, as in (1), is compared with the distribution of *no*-fragments, such as (2), with data from the two releases of the *British National Corpus* (BNC): BNC1994 (BNC Consortium 2007) and BNC2014 (Love et al. 2017; Brezina et al. 2021), available through #LancsBoxX (Brezina & Platt 2023). The main aim of the study is to determine whether *not*-(vs. *no*-)fragments replicate the trend of *not*-(vs. *no*-)negation towards informal language and speech described in, among others, Biber et al. (1999: 159) and Herrero- Zorita (2013).

(1) That's why! You've got to! *Not a problem*! I know. Derek! Now come on (BNC1994; IspKCP-21)
(2) Make up? *No problem.* Now you can chuck on a Snapchat filter: (BNC2014; NewSeSut729)

The findings reveal an increase in the use of *not*-fragments in speech and informal writing in recent diachrony (BNC1994–BNC2014). Specifically, the frequent *not*-fragments *not a chance* and *not a problem* tend to be avoided as fragmentary expressions in, for example, academic prose. These findings give support to the connection between *not*-fragments and colloquialisation, substantiated by the former's increasing occurrence in informal contexts.

References
Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan (eds.) 1999. *Longman grammar of spoken and written English*. Harlow: Longman.
BNC Consortium. 2007. British National Corpus: XML edition. Oxford Text Archive. http://www.natcorp.ox.ac.uk/cpr.xml?ID=reference (14 September, 2022).
Bowie, Jill & Bas Aarts. 2016. Clause fragments in English dialogue. In María José López-Couso, Belén Méndez-Naya, Paloma Núñez-Pertejo & Ignacio M. Palacios-Martínez (eds.) *Corpus linguistics on the move* (Language and Computers), vol. 79, 259–288. Leiden & Boston: BRILL. https://doi.org/10.1163/9789004321342_013
Brezina, Vaclav, Abi Hawtin & Tony McEnery. 2021. The written British National Corpus 2014 – design and comparability. *Text & Talk* 41(5–6). 595–615. https://doi.org/10.1515/text- 2020-0052
Brezina, Vaclav & William Platt. 2024. #LancsBox X. [Software v4.0.0]. Lancaster (UK): Lancaster University. https://lancsbox.lancs.ac.uk/
Cappelle, Bert. 2020. *Not on my watch* and similar *not*-fragments: Stored forms with pragmatic content. *Acta Linguistica Hafniensia* 52(2), 217–239. https://doi.org/10.1080/03740463.2020.1812365
Fernández-Pena, Yolanda. 2021. Towards an empirical characterisation and a corpus-driven taxonomy of fragments in written contemporary English. *Revista Electrónica de Lingüística Aplicada* 20(1), 136–154.
Herrero-Zorita, Carlos. 2013. A statistical study of the usage of *no*-negation and *not*-negation in spoken academic English. *Procedia - Social and Behavioral Sciences* 95, 482–489. https://doi.org/10.1016/j.sbspro.2013.10.672
Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316423530
Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & 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
Mair, Christian. 1997. The spread of the *going-to*-future in written English: A corpus-based investigation into language change in progress. In Raymond Hickey & Stanislaw Puppel (eds.) *Language history and linguistic modelling. A Festschrift for Jacek Fisiak on his 60th birthday*, 1537–1543. Berlin: De Gruyter.
Smitterberg, Erik. 2021. *Syntactic change in late Modern English: studies on colloquialization and densification*. Cambridge, UK: Cambridge University Press.

Oh yeah definitely: Exploring recent developments of the epistemic modal adverb *definitely*

Karin Aijmer (University of Gothenburg)

With the help of spoken corpora we can now study linguistic changes taking place in the 21st century. The changes involve frequencies and are both linguistic and sociolinguistic. The starting point for the present study is the observation based on comparable spoken corpora that the epistemic modal adverb *definitely* is increasing in frequency over a short period of time in present-day English while the related modal adverbs *certainly* and *surely* do not present a similar change (cf. Byloo et al. 2006; Downing 2001, Ranger 2011; Simon-Vandenbergen and Aijmer 2007). The aim is to describe the ongoing changes of *definitely* using spoken corpora from two different periods of time as my material. The research questions are:

- How is *definitely* changing in frequency and multifunctionality over a short period of time?
- What is the age and gender of the speakers initiating the changes and promoting the spread of *definitely*?
- What are the social and cultural motivations behind the changes (style and discourse practice)?

The occurrences of *definitely* have been retrieved from the (sample version of the) Spoken British National Corpus 2014 containing c. 5 million words (Love et al. 2017). A comparison is made with the cases of *definitely* with regard to their frequencies and usage in the demographic component of the British National Corpus 1994 which has roughly the same size. *Definitely* occurred 1,368 times in the Spoken BNC2014 and 503 times in BNC1994. Different functions of *definitely* such as emphasizer and response marker are identified and further research into its changes over time is carried out by investigating how *definitely* is used differently by speakers depending on their age and gender (see also Simon-Vandenbergen 2008).

Definitely can have a strong epistemic modal meaning establishing that something is true without any doubt ('total certainty'). The epistemic meaning can be weakened, in which case the emphasizing meaning of *definitely* is foregrounded. The emphasizing *definitely* typically co-occurs with other emphasizing markers (*most definitely, just definitely*) and it can be repeated for greater force. *Definitely* can also be used by the speaker to take up a position of superior knowledge in the conversation ('I definitely think', 'I definitely will') (cf. Dendale 2020). Before adjectives, *definitely* needs to be analysed as an intensifier (*definitely annoying, definitely weird).* Finally, *definitely* functions as a boosted response marker (*oh yeah definitely)* with the meaning of agreeing with a previous speaker or acceptance (as a response to a question in the preceding turn.)

The findings of samples of 200 words of *definitely* from the two corpora show that *definitely* increases in frequency in its expressive meanings as an emphasizer and as a boosted response marker in the Spoken BNC2014. It is also increasing in frequency in the function where it is used by speakers to position themselves as the authority of knowledge in the conversation.

Based on the sociolinguistic metadata about the speakers in the corpora, it is shown that *definitely* is more frequent in the spoken language of the younger age groups in both corpora. In BNC1994 the users of *definitely* are mostly male but in the Spoken BNC2014 *definitely* is used predominantly by women. Young female speakers boost their utterances by means of *definitely* rather than use the unmarked assertion or *certainly* to assert that something is true. They are not afraid of stating forceful opinions but strengthen their emotional involvement in the conversation and their identification with the social group by using *definitely.*

References

Byloo, P., R. Kastein and J. Nuyts. 2006. On *certainly* and *zeker*. *Belgian Journal of Linguistics* 20(1): 45-72.

Dendale, P. 2020. Are "modal adverbs" automatically modal markers? The case of French *certainement* with its epistemico-modal and its evidential use. *Anuari de Filologia. Estudis de Lingüística* 10: 39-76.

Downing, A. 2001. Surely you knew! *Surely* as a marker of evidentiality and stance. *Functions of Language* 8(2): 251-282.

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

Ranger, G. 2011. Surely not! Between certainty and disbelief. *Discours: Revue de Linguistique, Psycholinguistique et Informatique* 8: 3-22.

Simon-Vandenbergen, A.M. 2008. *Almost certainly* and *most definitely*. Degree modifiers and epistemic stance. *Journal of Pragmatics* 40: 1521-1542.

Simon-Vandenbergen, A.M. and K. Aijmer. 2007. *The semantic field of modal certainty. A corpus-based study of English adverbs*. Berlin: Mouton de Gruyter.

'I'd give anything not to have such a strong position on this matter': A corpus of Scottish opinions on assisted dying

Marc Alexander and James Balfour (University of Glasgow)

Work-In-Progress

The law in Scotland currently prevents dying people from asking for medical help to die. From late 2021 to 2022 a consultation took place on a proposed Assisted Dying for Terminally Ill Adults Bill to enable mentally competent, terminally ill adults to be provided at their request with assistance to end their life. This consultation received the highest number of responses to date for any Members Bill consultation in the Scottish Parliament, and its materials were made public in 2023 in advance of the Bill moving forward in 2024. From the consultation materials, we built a 2.1m word corpus of all 12,314 written responses received where the respondent consented to the availability of their answers.

We report in this paper on work in progress with regards to the discourse of this highly emotive and controversial topic. As can be expected, those who self-report as strongly opposed or in favour of assisted dying polarise in their keywords, reflecting common themes in their arguments. Supportive respondents, for instance, legitimise their position by appealing to the autonomy and free will of sufferers (e.g. *choice, my, right, wishes, required, decision, able,* and *suffer**, whereas those opposed emphasise factors which potentially undermine free will (*vulnerable, pressure, society, burden, lethal*). Indeed, the latter more frequently use labels which draws emphasis to the criminality of assisted dying or its external agency (*suicide, killing*). In collocates, too, *believe* co-occurs for supportive respondents with *required, right,* and *everyone* and for opposing respondents with *sanctity, Christian,* and *firmly*. In particular, the consultation asking for reasons why a respondent is supportive or opposed often elicits deeply personal narratives. 'Watch* my' occurs 300 times (usually followed by *die* or *suffer*), with the construction's major collocates including *slowly, agony, deteriorate, horrible, months, waste, starve, horrendous,* and *pain*. Such narratives frequently include references to close family members (e.g. *dad, friend, partner, husband*). For those in support of assisted dying, there are over 150 instances of the construction 'he/she wanted/asked/wished' as part of the rationale for their support. Finally, a common comparison in the corpus is with animals, with WMatrix's USAS category L2 (Living creatures: animals, birds, etc.) being a highly key semantic domain for supportive responses compared to the whole corpus, only after pronouns, E4.1- Sad (which includes *suffering*), X7+ Wanted (including *choice*), and categories relating to the concept of unnecessity. This animal category includes 329 uses of *animal/cat/dog/horse/pet* next to a negative in expressions such as 'we won't let a dog suffer, so why do we force humans?'. We will report on these and other interim results from the corpus, as well as analysing the difference in responses from professional bodies (from the Royal College of Physicians and the Scottish Partnership for Palliative Care to the Bishops Conference of Scotland and the Scottish Pagan Federation).

References

McArthur, Liam. 2022. *Proposed assisted dying for terminally ill adults bill: Consultation document.* Edinburgh: Scottish Parliament.

Rayson, Paul. 2009. *Wmatrix: A web-based corpus processing environment*. Lancaster: Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

A corpus-assisted discourse studies approach to meeting leadership style in FOMC deliberations

Gisle Andersen and Christian Langerfeld (NHH Norwegian School of Economics)

It is widely recognised that meetings are among the most central work-related events in corporations and organisations and that the meeting constitutes a distinct genre, distinguishable from other forms of Language for Specific Purposes (Angouri & Marra, 2010; Asmuß & Svennevig, 2009; Authors, 2023; Boden, 1994; Schwartzman, 1989; Svennevig, 2012; Tannen, 1994). Our paper expands upon earlier discourse-analytic work on meetings by investigating how meeting leadership style is manifested in the interactional behaviour of participants in high-stakes meetings. Specifically, we consider the ways in which the deliberations of the Federal Open Market Committee (FOMC) in the United States are characterised by humour and interruptions. We explore the recently compiled FOMC corpus (Langerfeld & Andersen, 2023) that contains approx. 11 million words of meetings transcripts from 1987 to 2017. During this period the FOMC had three different chairs. The FOMC is part of the US Federal Reserve system and makes key decisions on monetary policy that impact the American and global economies, such as the key interest rate. Methodologically, we base our study on a complementary quantitative and qualitative approach, as embodied in the framework of Corpus-Assisted Discourse Studies (Gillings et al., 2023).

Meetings are a key instrument for decision-making, and we can expect features of leadership style to have a bearing upon the ways leaders manage the meetings which they chair. We define meeting leadership style as the set of interactional features that a meeting chair deploys to exercise the leadership of meetings (Boden, 1994; Edelsky, 1993; Holmes & Stubbe, 2003). Our study aims at identifying potential differences between the three FOMC chairs who appear in the corpus (Greenspan, Bernanke, Yellen) as regards meeting leadership style, asking whether it is possible to identify consistency within each chair or adoption of different styles within the same chairship. We choose to focus on two readily accessible features of style, namely humour and interruptions. These discourse phenomena can be directly observed from transcriptions of laughter and cut-off sentences and overlapping speech. This enables the study of these features at the level of individual tokens and at an aggregate level across different sections of the corpus by statistical as well as qualitative methods.

Our findings show that humour is utilised commonly by chairs and delegates in ways which are similar to but also distinct from everyday conversational humour (Norrick, 2003) and from humour in the workplace more generally (Holmes & Stubbe, 2003). The study also shows that, although humour and interruptions are not necessarily correlating phenomena, there are observable differences in the meeting leadership styles performed by the three chairs in the FOMC corpus and differences in the interactive style across parts of their chairship. Moreover, the data shows that the variability of laughter and interruptions coincides in interesting ways with the point in time at which an utterance occurs during a meeting, and to some degree with external events (e.g. the financial crisis of 2007).

References

Angouri, J. & Marra, M. (2010). Corporate meetings as genre: a study of the role of the chair in corporate meeting talk. *Text & Talk - An Interdisciplinary Journal of Language, Discourse & Communication Studies* 30(6), 615–636. https://doi.org/10.1515/text.2010.030

Asmuß, B. & Svennevig, J. (2009). Meeting talk: An introduction. *Journal of Business Communication* 46(1), 3–22. https://doi.org/10.1177/0021943608326761

Langerfeld, C. & Andersen, G. (2023). The dynamics of turn-taking in meetings of the federal open market committee. *Fachsprache* 45(3-4) s, 187-210

Boden, D. (1994). *The business of talk: Organizations in action*. Polity Press.

Edelsky, C. (1993). Who's got the floor? In D. Tannen (Ed.), *Gender and conversational interaction* (pp. 189–227). Oxford University Press.
Gillings, M., Mautner, G. & Baker, P. (2023). *Corpus-assisted discourse studies*. *Elements in corpus linguistics*. Cambridge University Press. https://doi.org/10.1017/9781009168144
Holmes, J. & Stubbe, M. (2003). *Power and politeness in the workplace*. Pearson.
Norrick, N. R. (2003). Issues in conversational joking. *Journal of Pragmatics* 35(9), 1333–1359. https://doi.org/10.1016/S0378-2166(02)00180-7
Schwartzman, H. (1989). *The meeting: Gatherings in organizations and communities*. Plenum Press.
Svennevig, J. (2012). Interaction in workplace meetings. *Discourse Studies* 14(1), 3–10. https://doi.org/10.1177/1461445611427203
Tannen, D. (1994). Interpreting interruption in conversation. In D. Tannen (Ed.), *Gender and discourse* (pp. 53–83). Oxford University Press.

Register and social stratification in a new corpus of Bislama, an English-based creole

Carol Aru[1], Jocelyn Aznar[2], Manfred Krifka[2], Miriam Meyerhoff[3] and Tonjes Veenstra[2]

([1]Humboldt University of Berlin, [2]Leibniz ZAS Berlin, [3]University of Oxford)

Bislama is the English-based creole spoken widely in Vanuatu (SW Pacific). It is the national language of the country and it is spoken by nearly all of the c.300,000 Ni-Vanuatu. It is the first language of about 10% of the population (Vanuatu 2020). Our team has compiled the first age- and sex-balanced corpus of spoken Bislama as used by residents of the capital, Port Vila. The corpus consists of 41 speakers (three age groups), nearly 18 hours of speech with over 122,000 words of Bislama (transcribed and translated). Some speakers have been recorded multiple times with different addressees, speaking on different topics. This allows us to explore possible register/style effects in Bislama. Our paper introduces the corpus to the ICAME community and examines variation across four variables.

Two of the variables are phonetic: phonetic reduction of auxiliary verbs (1); final consonant reduction/cliticisation of prepositions (2). Two are morpho-syntactic: plural marking on NPs (3); plural subject agreement on verbs (4).

(1) a. Yumi **stap** ([stap]) praktisim lanwis blong yumi. 'We keep using our native languages'
    b. Jaena i **stap** ([sta]) mekem blo hem naoia. 'China's making its move now'.
(2) a. Mi stap helpem faenans **blong olgeta** ([blɒʔɒlgɛta]). 'I'm helping them out with their finances'.
    b. Taem i tanem i go **long olgeta** ([lɒlgɛta])… 'When he turns to them…'
(3) a. **Sam brata** o sista oli maret. 'Some of my brothers and sisters are married'.
    b. **Sam** narafala **mats** oli gat mak tu. 'Some other mats also have markings'.
(4) a. Ol pikinini bae **oli** stap kam tumoro. 'The children will be arriving tomorrow'.
    b. Ol materiol ia bae **i** olsem wanem. 'What will the materials be like?'

We find limited inter- and intra-speaker variation in morphosyntactic variables, more in phonological ones. Auxiliary reduction is strongly constrained by auxiliary type: *stap* (imperfective) is realised as [sta] (N=1044, 71%) more than it is as [stap]; whereas *save* (ability) is realised as [savɛ] (N=395, 86%) more than it is as [sae]. Prepositions before pronouns are usually cliticised (N=1077, 82%).

In previous work, we have shown that variation in NP plural marking (N=2308) is primarily constrained by topic – non-canonical, innovative forms of the plural occur most in talk about work, then in conversations and oral histories, never in traditional stories. Agreement with 3rd person plural subjects shows little variation (85% canonical agreement [4a], N=981). More detailed analysis of the plurals showed that syntactic factors were significant, indicating that this variable is, indeed, morphosyntactic. It appears that in Bislama (like many varieties of English, Labov 1993, Smith et al. 2013) morphosyntax lies below the level of conscious awareness and is less amenable to being recruited for social/symbolic differentiation of speakers in the speech community. Our results suggest that members of the Bislama speech community of Port Vila share the same grammar, but differentiate themselves from each other phonologically or phonetically.

Cross-corpora diversity in web and spoken data: *BE going to* and its variants (*gonna*, *imma*, etc.)

Leela Azorin (University of Aix-Marseille)

Work-In-Progress

This presentation will focus on the study of the emergent semi-modal (Collins 2009, Machová 2015) *BE going to* and its variants (*gonna, gon, gunna, imma,* etc.) in two corpora: a web corpus (Climate Change Tweets, Littman and Wrubel 2019) and a spoken corpus with naturally occurring interactions (Santa Barbara Corpus (SBC), Du Bois et al. 2000-2005). *Gonna* is usually seen as the last step in the grammaticalization process of *BE going to* (Traugott and Trousdale 2013) but it still suffers from a lack of description, as it is most often seen as "the informal variant" of *BE going to* (Berglund 2000) and only mentioned in relation to *BE going to* in traditional grammars, as pointed out by Col and Duchet (2001). The two chosen corpora highlight the importance of *gonna* – notably in spoken data – and other variants outside *BE going to*, questioning the place and categorization of such variants within English grammar (Lorenz 2013, Daugs 2021). Are they mere pronunciation variants or morphosyntactic variants in their own right?

The web and spoken data investigated shed light on the variation under scrutiny and on the more innovative forms of the paradigm, showing that this paradigm cannot be reduced to the contrast between *BE going to* and *gonna* since it exhibits other forms (such as *imma* or *gon*). These more innovative forms have mostly gone unnoticed in more traditional corpora such as the BNC or the COCA (Davies 2008-) for the British and American dialects. Our research thus advocates the use of more diverse material as it can act as a relevant heuristic tool.

During this presentation, we will first set up a typology of the different variants found in the corpora and examine whether these variants are the same for the written web corpus and the spoken corpus, comparing the two types of data. Then, we will try to account for this diversity of forms. To do so, we investigate several criteria, which are being compared cross-corpora:

- The presence or absence of *BE*
- The presence or absence of a pronoun
- The type of verb following the variants
- The presence of an adverb
- The pragmatic discursive function of the variant

These criteria were selected after a preliminary analysis of all the variants, noting that the copula or a subject may be absent from the sentence including the semi-modal.

Our preliminary results show that:

- There is a diversity of variants inherent in the *BE going to/gonna* paradigm*: goin to, gon, gunna, imena, imma,* etc.
- The absence of the copula seems to be correlated with a more contracted variant than *going to* in both corpora.
- There exist syntactic criteria constrain variation, such as the pronoun *I* in *imena* and *im(m)a*.
- Moreover, the data highlight a potential discursive function of *gonna*, especially in the spoken corpus, in sentences such as *I was gonna say* (SBC 51: 1483); *What I was gonna ask* (SBC 52: 1439).

Therefore, it seems that using and cross-comparing different corpora can be useful to put forward variants or pragmatic functions of forms that may not have been prominent in more traditional corpora, although these variants and functions do exist, and they are essential if one wants to study a paradigm in its entirety and in its diversity.

References

Berglund, Ylva. 2000. *Gonna* and *Going to* in the Spoken Component of the British National Corpus. In *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerised Corpora (ICAME20),* edited by Christian Mair and Marianne Hundt, 35-49. Brill. https://doi.org/10.1163/9789004490758_005

The British National Corpus, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Col, Gilles and Jean-Louis Duchet. 2001. Forme non stable et grammaticalisation: le cas de gonna en anglais contemporain. In *Grammaticalisation 2*. *Concepts et cas. Travaux linguistiques du CERLICO* 14. Rennes: Presses universitaires de Rennes.

Collins, Peter. 2009. *Modals and Quasi-Modals in English*. Rodopi.

Daugs, Robert. 2021. Contractions, Constructions and Constructional Change: Investigating the Constructionhood of English Modal Contractions from a Diachronic Perspective. In *Modality and Diachronic Construction Grammar*, edited by Martin Hilpert, Bert Cappelle and Ilse Depraetere, 13–52. Constructional Approaches to Language 32. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.32.02dau

Davies, Mark. 2008-. The Corpus of Contemporary American English (COCA). Available Online at https://www.English-Corpora.Org/Coca/.

Du Bois, John, et al. 2000. *Santa Barbara Corpus of Spoken American English*. Vol. Parts 1-4. Linguistic Data Consortium. Philadelphia. https://www.linguistics.ucsb.edu/research/santa-barbara-corpus#Contents.

Littman, Justin and Laura Wrubel. 2019. Climate Change Tweets Ids. Harvard Dataverse. https://doi.org/10.7910/DVN/5QCCUU

Lorenz, David. 2013. From Reduction to Emancipation: Is Gonna a Word? In *Studies in Corpus Linguistics*, edited by Hilde Hasselgård, Jarle Ebeling and Signe Oksefjell Ebeling, 133–52. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.57.11lor

Machová, Dagmar. 2015. The Degree of Grammaticalization of Gotta, Gonna, Wanna and Better: A Corpus Study. *Topics in Linguistics* 15 (1). https://doi.org/10.2478/topling-2015-0005

Traugott, Elizabeth Closs and Graeme Trousdale. 2013. *Constructionalization and Constructional Changes*. Oxford Studies in Diachronic and Historical Linguistics. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199679898.001.0001

Characterising religious vocabulary in sixteenth-century Scotland

Beth Beattie (University of Glasgow)

The idea that different ideologies can be distinguished by their vocabulary is well-established (Williams, 2015), and there is a small body of research that has applied this principle to historical religious groups: Hudson (1981) explored the linguistic usages of the Lollards, and "godly" has long-since been associated with Protestants and other evangelical denominations (Collinson, 1983; Smith, 2020). Developments in corpus analysis techniques have resulted in more systematic descriptions of the discourses of Christian denominations in sixteenth-century England (Smith, 2021). However, following the Reformation, the religious makeup of Scotland was markedly different from that of England, but there are ideological similarities between Catholic- and Protestant-aligned denominations (Williamson, 2022). Furthermore, the ties between linguistic and socio-cultural practices in religious domains form an important part of exploring broader connections between and across discourse communities, which also contributes to evolving British national identities. This paper addresses the extent to which models of English religious discourse can be applied to the Scottish context in the same period and, if not, what adjustments are required to create a description of Reformation era Scottish religious discourse.

This research necessitated the creation of a 350,000-word corpus of sixteenth-century Scottish religious texts, derived from Early English Books Online and the National Library of Scotland. The corpus includes works by prominent Scottish religious figures like John Knox and Quintin Kennedy and is divided into two subcorpora – "Protestant" and "Catholic" – with both subcorpora containing texts written in Early Modern English and Older Scots. This research takes the vocabulary lists established by Smith (2020) of lexemes coded as "evangelical" and "Roman Catholic" in English religious discourse and applies them to this corpus, with the Dictionary of the Scots Language being used to identify Scots equivalents for English lexemes when needed. Frequency and keyness analysis is used to identify the most salient lexemes for each subcorpus, with collocation and concordance analysis providing further detail about the contexts for each lexeme.

Results indicate that there is a greater similarity between Catholic vocabulary found in Scotland and England than evangelical, with the majority of Catholic-coded lexemes being identified as keywords in the Catholic subcorpus but not in the Protestant subcorpus. However, the same cannot be applied to the evangelical-coded lexemes. Scottish Catholic writers are found to have used some of these words to the same degree as the Scottish Protestants, with "godly" being identified as a keyword in both subcorpora. Furthermore, the similarity of the contexts in which "godly" and other words are found in both subcorpora indicates that Scottish Catholics are deliberately emulating Protestant discourse, possibly in an attempt to make their ideology more palatable in a post-Reformation Scotland (see Ryrie, 2004). These results indicate that while there are similarities between English and Scottish religious discourses, they are denomination-dependent. Therefore, a Scottish framework needs to acknowledge the socio-cultural and religious variation not accounted for in an English framework.

References

Collinson, P. (1983) *Godly People: Essays on English Protestantism and Puritanism*. London: The Hambledon Press (History Series, 23).

Hudson, A. (1981) A Lollard sect vocabulary?, in M. Benskin and M.L. Samuels (eds.) *So Meny People Longages and Tonges: Philological Essays in Scots and Mediaeval English Presented to Angus McIntosh*. Edinburgh: Middle English Dialect Project, pp. 15–30.

Ryrie, A. (2004) Reform without frontiers in the last years of Catholic Scotland. *The English Historical Review* 119 (480), 27–56.

Smith, J.J. (2020) Godly vocabulary in Early Modern English religious debate, in E. Jonsson and T. Larsson (eds.) *Voices Past and Present - Studies of Involved, Speech-related and Spoken Texts: In Honor of Merja Kytö*. Amsterdam: John Benjamins Publishing (Studies in Corpus Linguistics, 97), pp. 95–112.
Smith, J.J. (2021) Lexical choices in Early Modern English devotional prose. *Journal of Historical Pragmatics* 22(2), 263–281.
Williams, R. (2015) *Keywords: A Vocabulary of Culture and Society*. Third Edition. New York: Oxford University Press.
Williamson, A.H. (2022) Britain Reformed: Competing Visions, 1527-1641, in W.I.P. Hazlett (ed.) *A Companion to the Reformation in Scotland, ca. 1525-1638: Frameworks of Change and Development*. Leiden, Boston: Brill (Brill's Companions to the Christian Tradition), pp. 660–688.

Postnominal adjectives in Old English revisited: A reductive comparative approach

Kristin Bech and Tine Breban (University of Oslo, The University of Manchester)

Prenominal position is the default position for attributive adjectives in Old English, but there are exceptions: in the *York–Toronto–Helsinki Parsed Corpus of Old English Prose* (YCOE, Taylor et al. 2003), 4.1% of adjectives occur postnominally (1,821 out of 44,196). The factors influencing the position of adjectives in Old English have been the subject of multiple studies (e.g. Bech 2017, 2019; Bech et al. 2024; Fischer 2000, 2001; Grabski 2017, 2020; Haumann 2010; Pysz 2009; Sampson 2010). However, variation in adnominal position was no longer a fully productive system in Old English, which makes it difficult to gauge the factors conditioning the variation and the role they played. Consequently, some studies have relied on illustrative examples to support their analyses, and others found limited significance for any of the factors included in a regression analysis. Many discourse-related and semantic factors are moreover notoriously difficult to annotate, e.g. stage versus individual level meaning, given versus new, which is a concern for the replicability and reliability of any (large-scale) studies.

The question we ask in this paper is whether a different way of approaching the same data could break this stalemate. We use what we call a reductive comparative approach to argue that most instances of postnominal adjectives in Old English can be explained by straightforwardly operationalizable factors which are found to play a role in the variation in other early Germanic languages (cf. Bech et al. 2024). We start from all examples of postnominal adjectives in the YCOE corpus, and gradually eliminate examples that can be explained by factors which played a role to varying degrees in Germanic languages in general, such as lexically motivated patterns, weight of the adjective phrase, and weight distribution with regard to the head (Bech et al. 2024).

In this way we step by step remove (1) postnominal adjective phrases containing more than a single item (367 examples); (2) postnominal tokens that represent frequently recurring adjective types, e.g. *self*, *almighty* in the phrase *God ælmihti*, *full* in *anne cuculere fulne* 'a spoon full', which could be considered lexically motivated (1,217 examples); (3) noun phrases that contain both a prenominal and a postnominal adjective ('flanked adjectives') (53 examples); (4) noun phrases that contain a prenominal quantifier or possessive marker and a postnominal adjective (cf. Spamer 1979; Sampson 2010) (58 examples).

At the end of this process, only 63 or 3.5% of the original sample of postnominal adjectives still need explaining. Of these 24 examples are found in medical texts (Bald's Leechbook, Lacnunga, Herbarium). Taking into account Grabski's (2017) finding that Latin influence did not play a significant role, we consider this a possible further indication of genre-specific productivity

of the postnominal position for adjectives in Old English. We conclude that due to the limited productivity of the postnominal placement, and the significant amount of data that can be explained with reference to weight, discourse-related and semantic factors are unlikely to be primary motivations for the postnominal placement of adjectives in Old English.

References

Bech, K. 2017. Old English and Old Norwegian noun phrases with two attributive adjectives. *Bergen Language and Linguistics Studies* (BeLLS) 8, 1–18. https://bells.uib.no/index.php/bells/article/view/1326.

Bech, K. 2019. Contextualizing Old English noun phrases. In K. Bech & R. Möhlig-Falke (eds.), *Grammar, discourse, context: Grammar and usage in language variation and change*, 15–48. Berlin, Boston: De Gruyter

Bech, K., H. Booth, K. Börjars, T. Breban, S. Petrova & G. Walkden. 2024. Noun phrase modifiers in early Germanic: A comparative corpus study of Old English, Old High German, Old Saxon, and Old Icelandic. In K. Bech & A. Pfaff (eds.), *Noun phrases in early Germanic languages.* Berlin: Language Science Press.

Fischer, O. 2000. The position of the adjective in Old English. In R. Bermúdez-Otero, D. Denison, R.M. Hogg & C.B. McCully (eds.), *Generative Theory and Corpus Studies. A Dialogue from 10 ICEHL*, 153–181. Berlin/New York: Mouton de Gruyter.

Fischer, O. 2001. The position of the adjective in (old) English from an iconic perspective. In O. Fischer & M. Nänny (eds.), *The Motivated Sign. Iconicity in Language and Literature* 2, 249–276. Amsterdam: John Benjamins.

Grabski, M. 2017. *The position of the adjective in Old English prose. A corpus study.* Doctoral thesis, University of Łódź.

Grabski, M. 2020. Three types of Old English adjectival postposition: A corpus-based Construction Grammar approach. *Journal of English Linguistics* 48(2), 166–198.

Haumann, D. 2010. Adnominal adjectives in Old English. *English Language and Linguistics* 14(1), 53–81.

Pysz, A. 2009. *The syntax of prenominal and postnominal adjectives in Old English*. Newcastle: Cambridge Scholars Publishing.

Sampson, S.A. 2010. Noun phrase word order variation in Old English verse and prose. PhD dissertation, Ohio State University.

Spamer, J.B. 1979. The development of the definite article in English: A case study of syntactic change. *Glossa* 13, 241–250.

Taylor, A., A. Warner, S. Pintzuk & F. Beths. 2003. *The York–Toronto–Helsinki Parsed Corpus of Old English Prose*. http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm.

Expanding the scope of grammatical variation: Towards a comprehensive account of genitive variation across registers

Doug Biber, Randi Reppen and Tove Larsson (Northern Arizona University)

Most previous studies of genitive variation in English have considered only the choice of two variants (*'s* versus *of*), based on analysis of only tokens that are judged to be interchangeable. In the present study (based on Biber et al. 2023; see also Szmrecsanyi et al. 2016), we explore the possibility of extending the scope of analysis in both respects, investigating the following research questions:

1) Can we account for the use of pre-modifying nouns as a third genitive variant?
2) Can we account for <u>all</u> tokens of the genitive construction in running discourse?

In addition, we extend the scope of analysis by exploring a larger question:

3) Do contextual constraints have different importance in different registers?

The first stage of analysis employs a text-linguistic register approach to compare the rates of genitive variants in conversation, newspaper reports, academic articles. The analysis shows that genitives overall are much more frequent in the written registers, with the pre-modifying noun variant being the most common. The *'s*-genitive is by far most common in news reportage. In conversation, the *'s*-genitive is infrequent and surprisingly accounts for only 7% of the total genitives. In academic writing, the *'s*-genitive accounts for only 2% of all genitives, but it actually occurs in texts more frequently than in conversation. NN genitives are especially common in written academic texts. Proportionally, NN constructions account for the majority of genitive tokens in all three registers.

Then, the second stage of analysis was to undertake a variationist analysis, to account for the choice of genitive variant in particular contexts and registers. 3,471 genitive tokens were hand-coded for 10 contextual characteristics (e.g., length of the Modifying NP, semantic category of the Modifying noun and the Head noun, final sibilancy of the Modifying noun). Statistical analyses with Random Forests and Conditional Inference Trees are triangulated, showing how contextual factors interact in predicting the use of each genitive variant – and how patterns of variation differ across registers.

The results show that the linguistic patterns of genitive variation – including both interchangeable as well as non-interchangeable (but non-categorical) tokens – can be accounted for with a high degree of accuracy. Two major contextual factors are especially important: the semantic category of the modifying noun, and the length of the modifying NP. In general, the *'s*-genitive is strongly associated with animate modifying nouns, and the *of*-genitive is strongly associated with long modifying NPs. However, beyond those two general trends, the results show a complex network of interacting factors associated with one or another of the variants, with different linguistic patterns of variation in each register.

References

Biber, D., B. Szmrecsanyi, R. Reppen and T. Larsson. 2023. Expanding the scope of grammatical variation: Towards a comprehensive account of genitive variation across registers. *English Language and Linguistics* 28, 95-133.

Szmrecsanyi, Benedikt, D. Biber, J. Egbert & K. Franco. 2016. Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change* 28, 1-29.

Large and tidy? A method for finding structure in mega-corpora

Axel Bohmann (University of Freiburg)

Register (Biber & Egbert 2023) is a key factor of text-linguistic differentiation (Biber 1989; Bohmann 2020). At the same time, it is largely absent as a category in the sampling frames of many large-scale corpora. Responses to the launch of the Corpus of Global Web-based English (GloWbE; Davies & Fuchs 2015 2015), for instance, have expressed concern about the lack of genre information in this data base (Mair 2015; Mukherjee 2015; Nelson 2015).

The present study explores an innovative method for imputing register structure in large, genre-lean corpora. It does so by combining small and carefully curated corpora with big data in one multi-dimensional analysis, specifically: corpora from the International Corpus of English (ICE) project (Greenbaum & Nelson 1996) and GloWbE. Frequency information for 150 linguistic features is extracted from each of the 1,771,698 texts in this combined corpus, and dimensions of variation are established based on factor analysis of the resultant text-feature matrix. The advantage of this method is that it derives dimensional information primarily based on the (numerically dominant) texts in GloWbE while being immediately interpretable in relation to the

ICE corpora with their detailed text-category information. This, in turn, allows for the interpretation of GloWbE texts with reference to the space of register variation covered in ICE.

Four dimensions are established in this procedure, which correspond to interpretable textual differences: 1) verbal versus nominal discourse, 2) informational versus argumentative focus, 3) addressee-orientation, and 4) narrative. While these dimensions are primarily derived on the basis of variation in GloWbE (which contributes 98.5% of all texts in the combined corpus), they are not strongly correlated with the corpus' internal distinction between blog and website data. This fact highlights the need to more actively attend to the corpus' heterogeneity in terms of register.

In contrast to extant work (e.g., Biber & Egbert 2018), this approach does not aim to identify register labels for corpus texts, but to situate each text in an interpretable space of (register-based) linguistic variation. This is consistent with Biber et al.'s (2020) treatment of register in continuous terms.

In addition to presenting the methodological details and results of this general procedure, this talk also showcases the method's utility through a case study of variation in GloWbE. An analysis of variation in future-time reference between *will* and BE *going to* shows that register information derived in the steps outlined above significantly improves the quality of statistical models in single-feature studies.

References

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

Biber, Douglas and Jesse Egbert (2018). *Register Variation Online*. Cambridge: Cambridge University Press.

Biber, Douglas and Jesse Egbert (2023). What is a register?: Accounting for linguistic and situational variation within–and outside of–textual varieties. *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.

Bohmann, Axel (2020). *Variation in English Worldwide: Registers and Global Varieties* (Studies in English Language). Cambridge: Cambridge University Press.

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.

Greenbaum, Sidney and Gerald Nelson (1996). The International Corpus of English (ICE) Project. *World Englishes* 15 (1): 3–15.

Mair, Christian (2015). Response to Davies and Fuchs. *English World-Wide* 36 (1): 29–33.

Mukherjee, Joybrato (2015). Response to Davies and Fuchs. *English World-Wide* 36 (1): 34–37.

Nelson, Gerald (2015). Response to Davies and Fuchs. *English World-Wide* 36 (1): 38–40.

Introducing the Corpus of Young German Learner English

Lea Bracke[1], Robert Fuchs[2], Anna Rosen[3], Bethany Stoddard[4] and Valentin Werner[1]

([1]University of Bamberg, [2]University of Hamburg, [3]University of Freiburg, [4]University of Bonn)

Work-In-Progress

Although Learner Corpus Research (LCR) has contributed significantly to a better understanding of Second Language Acquisition (SLA) processes in general, its full potential for the analysis of interlanguage (Selinker 1972, 1992) is yet to be realized. This is due to several major challenges identified in the literature (e.g. Myles 2021; Tracy-Ventura et al. 2021). These include, among others, (i) an underrepresentation of beginner and lower-intermediate learners, (ii) an underrepresentation of spoken material and truly bi-modal data (i.e. data in different modes

produced by the same learners), (iii) a lack of or unsystematic elicitation of metadata, (iv) a lack of longitudinal or at least quasi-longitudinal perspectives, and (v) a neglect of task effects.

The project presented in this poster will address these challenges by compiling and analyzing a corpus of Young German Learner English (YGLE). English is taught as a compulsory (mostly, first) foreign language to the vast majority of secondary schools students in Germany. Despite the important role that the teaching of English plays in the German education system, relatively little representative information is available on the overall learning outcomes, common learner errors and learning trajectories. This research gap can be addressed by LCR, which has the potential to provide representative and reliable information on the described target group of learners (Mukherjee 2008).

The YGLE corpus project thus aims to complement the extensive body of work on highly advanced L1 German EFL learners (e.g. Fuchs et al. 2016; Römer et al. 2020), based on various corpora of learners at the university level, by creating a database on the production of beginning to intermediate L1 German EFL learners in institutional contexts.

To this end, data are being collected from around 700 participants at secondary schools (learners aged 10–18 years) across the German three-tier school system. To represent a wide range of communicative contexts, the task types administered include both established (timed argumentative essay, picture description) and innovative task types (group discussion, elicitation of digital communication) with varying degrees of planning and interactivity. In addition, an extensive set of metadata is collected, based on a modified version of the questionnaire and procedure proposed by Möller (2017) and in line with the core L2 metadata scheme (Frey et al. 2023). This metadata comprises established and validated test batteries assessing information on socioeconomic and educational status, linguistic background, language use across different social contexts (including exposure to English outside of school), motivation (standardized tests FLM 3–6 R, FLM 7–13; Lohbeck & Petermann 2019; Petermann & Winkel 2015), as well as general and verbal cognitive abilities (standardized test AID-G; Kubinger & Hagenmüller 2019).

After transcription and annotation with a focus on items relevant for the complexity-accuracy-fluency (CAF) triad, interactions between the CAF components as well as the influence of contextual and learner variables will be assessed using mixed-effects regression modeling. YGLE will eventually be made available to the LCR community, allowing (i) the exploration of areas beyond CAF (e.g. phonology, learner pragmatics, etc.) and potentially (ii) comparison with data from beginner and intermediate learners of English worldwide.

References

Frey, J. C., König, A., Stemle, E. W. & Paquot, M. (2023). A core metadata schema for L2 data. Paper presented at *EuroSLA 32, Conference of the European Second Language Association*. August 2023, Brimingham.

Fuchs, R., Götz, S. & Werner, V. (2016). The present perfect in learner Englishes: A corpus-based case study on L1 German intermediate and advanced speech and writing. In V. Werner, E. Seoane & C. Suárez-Gómez (Eds.), *Re-Assessing the Present Perfect* (pp. 297–337). Mouton de Gruyter.

Kubinger, K. & Hagenmüller, B. (2019). *Gruppentest zur Erfassung der Intelligenz auf Basis des AID*. Hogrefe.

Lohbeck, A. & Petermann, F. (2019). *Fragebogen zur Leistungsmotivation für Schülerinnen und Schüler der 3. bis 6. Klasse – Revision*. Hogrefe.

Möller, V. (2017). *Language Acquisition in CLIL and Non-CLIL Settings: Learner Corpus and Experimental Evidence on Passive Constructions.* Benjamins.

Myles, F. (2021). Commentary: An SLA perspective on learner corpus research. In B. Le Bruyn & M. Paquot (Eds.), *Learner Corpus Research Meets Second Language Acquisition* (pp. 258–273). Cambridge University Press.

Petermann, F. & Winkel, S. (2015). *Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13. Klasse.* Pearson Harcourt.

Römer, U., Salicky, S. C. & Ellis, N. C. (2020). Verb-argument constructions in advanced L2 English learner production: Insights from corpora and verbal fluency tasks. *Corpus Linguistics and Linguistic Theory* 16(2), 303–331.
Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(1/4), 209–232.
Selinker, L. (1992). *Rediscovering Interlanguage*. Longman.
Tracy-Ventura, N., Paquot, M. & Myles, F. (2021). The future of corpora in SLA. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 409–424). Routledge.

Paradigmatic enrichment of constructional paradigms: A diachronic perspective

Lieselotte Brems (KU Leuven)

This paper is a diachronic follow-up of Brems (2022) and looks into the size noun expression *bunch of* on the basis of extractions made from COHA (*Corpus of Historical American English*) for each decade of the available time spans. It is interested in comparing patterns where the size noun expression is preceded by a premodifier, so as to substantiate the claim made in Brems (2022) that size noun expressions enrich the quantifier paradigm by allowing (restricted) premodification, rather than seeing this as incomplete decategorialization (Hopper 1991) and absence of grammaticalization. This study hence draws attention to the underresearched but fundamental aspect of grammaticalization, namely what Diewald & Smirnova (2010) call the paradigmatic phase. Studying this paradigmatic phase is crucial to understanding how grammar and grammaticalization work; yet, it has been grossly overlooked in grammaticalization research even though it quite recently seems to be on the linguistic agenda again.

The corpus data show that *bunch of* has been premodified by adjectives that are both qualitative and quantitative in nature, without affecting the grammatical status of the size noun expression, and that size nouns in general form semi-stable subparadigms within the paradigm of quantifiers, which they enrich by allowing for more fine-grained quantitative information.

I will look at paradigms as constructional networks with different levels of schematicity, micro-, meso- and macro-level. For size nouns, the macro-level concerns the general function of quantification with regard to which size noun expressions are a meso-construction built on NP of NP syntagms. Each size noun counts as a micro-construction. Complex subordinators are subparadigms, or meso-constructions, within the paradigm of subordinators, with specific complex subordinators again functioning as micro-constructions, each displaying their own behaviour, collocational preferences and degrees of paradigmatic enrichment.

With these case studies, I zoom in on what happens in and 'after' grammaticalization, as expressions settle into a grammatical paradigm. How do specific paradigms' internal dynamics work? How are relations between potentially competing members of one paradigm (re)defined and how does a division of labour come about? I will argue that in the case studies at hand, within existing paradigms, periphrastic subsystems are integrated that are productive and semi-stable systems.

References

Brems, Lieselotte. 2022. Semi-stable systems in PDE: Paradigmatic enrichment of constructional paradigms. Paper presented at ICAME43.
Collins Wordbanksonline: https://wordbanks.harpercollins.co.uk/
Hopper, Paul J. 1991. On some principles of grammaticization. In Elizabeth C. Traugott & Bernd Heine (eds.) *Approaches to Grammaticalization, Volume 1*. Amsterdam: John Benjamins, 17-36.

Modeling morphosyntactic variation in Sint Maarten English(es)

Sarah Buschfeld and Andreas Weilinghoff (TU Dortmund, University of Koblenz)

Due to a complex history of colonization, foreign administration, and migration, Sint Maarten, the southern part of the Eastern Caribbean island of St. Martin, is a highly multilingual territory. While Dutch is the colonial language of this part and still widely spoken and used, in particular in the administrative domains, English is the L1 for the majority of Sint Maarteners. Even though many parts of the Caribbean, and in particular the many Englishes and English-based creoles spoken there, have been extensively studied, the status and linguistic characteristics of the English(es) spoken in Sint Maarten have not yet been investigated in a comprehensive way.

The paper seeks to address this research gap and is a first step towards inquiring into the following research questions: What are the features of Sint Maarten English? How is it different from or similar to other Caribbean Englishes/English-based creoles? To this end, we exemplarily investigate two morphosyntactic features, viz. the realization of past tense and progressive marking, both of which have been reported to have local realizations in Eastern Caribbean Englishes/Creoles (e.g. Aceto 2008). The data come from Labovian-style sociolinguistic interviews with 35 Sint Maarteners and from three hours of free interaction between locals, without a researcher present. The data were coded for local vs. BrE/AmE realizations of the features under observation. The influence of independent variables on the frequencies of these forms (e.g. age, gender, ethnicity, and intralinguistic predictors) is modeled by means of repeated undersampling in ctrees to meet the imbalance between local and BrE/AmE speech forms (cf. the PrInDT approach, Weihs & Buschfeld 2021a, b). The analysis also incorporates regression modeling which not only provides another statistical perspective on the observed patterns but also allows for a comparison between the two approaches (ctress and regression analysis).

The findings suggest that Sint Maarteners use BrE/AmE, creole-like, and other local realizations of the features, influenced by the extralinguistic variables under investigation. Moreover, the results also show that Sint Maarten English clearly distinguishes itself from other anglophone creoles in the Caribbean. We finally discuss what these findings imply for the long-assumed dichotomy between varieties of English and their local realizations on the one hand and English-based creoles on the other. Especially within the context of the World Englishes paradigm, Sint Maarten English stands out a unique case due to the influence of both French and Dutch colonization on the island.

References

Aceto, Michael. 2008. Eastern Caribbean English-derived language varieties: Morphology and syntax. In Edgar W. Schneider (ed.), *The Americas and the Caribbean*. Berlin: De Gruyter, 645–660.

Weihs, Claus & Buschfeld, Sarah. 2021a. Combining prediction and interpretation in decision trees (PrInDT) - a linguistic example, available at arXiv: http://arxiv.org/abs/2103.02336

Weihs, Claus & Buschfeld, Sarah. 2021b. Repeated undersampling in PrInDT (RePrInDT): Variation in undersampling and prediction, and ranking of predictors in ensembles, available at arXiv: https://arxiv.org/abs/2108.05129.

Modelling argument omission: A cognitive and multivariate study of object null instantiation in English(es)

Vladimir Buskin (KU Eichstätt-Ingolstadt)

A large, yet not clearly delineated group of transitive verbs are known to license object omission in English. The implicit objects are associated with several distinct interpretive strategies, which include anaphoric (cf. 1) and indefinite (cf. 2) construals, among others. In the Construction Grammar literature, these modes of argument omission are referred to as Definite Null Instantiation (DNI) and Indefinite Null Instantiation (INI), respectively (Fillmore and Kay 1995: §7-3ff). While these phenemona have received some attention in the past two decades (Lambrecht and Lemoine 2005; Lyngfelt 2012; Ruppenhofer and Michaelis 2010), there is as of yet only a small body of theoretical research and a surprsing lack of rigorous quantitative analyses.

(1) *I won* $\emptyset_{\text{DNI:Game}}$.
(2) *I haven't eaten* $\emptyset_{\text{INI:Food}}$ *yet.*

The exact reasons for the availability of null objects continue to be a matter of much dispute; however, besides discourse-pragmatic aspects, the semantic properties of the verb are repeatedly highlighted as potentially influential. These include iterativity, genericity and to a lesser extent telicity (Goldberg 2001), which may interact with the constructions in which the verb occurs (Lemmens 2006); Ruppenhofer (2004) adduces the semantic frame evoked by the verb as a further variable. More generally, argument omission raises the question of what areas of linguistic knowledge are required to produce and process elliptic utterances. Present accounts vary from null instantiation as a property of lexemes (Fillmore 2007: 146) or constructions (Croft 2001: 277) to a valency-reducing mechanism (Ruppenhofer and Michaelis 2010: 175).

Assuming a Usage-based Construction Grammar framework (Diessel 2019, 2020; Hoffmann 2022), the primary goal of this study is to develop cognitively plausible models of argument omission based on naturally occurring language data. In fact, it is claimed that null instantiation phenomena form a family of abstract, yet semantically rich constructions, i.e., complex Saussurean signs that are stored in the mental language network. Their distributional properties will be established by means of a comprehensive corpus study of verb complementation patterns in L1- and L2-Englishes.

In terms of method, approx. 200 verbs currently known to exhibit said alternation (see Herbst et al. 2004; Levin 1993) are examined in several national components of the International Corpus of English, which takes into account a broad range of spoken and written registers. As soon as the annotation process has been completed, the data will be subjected to unsupervised and supervised machine-learning algorithms to perform cluster analyses (principal component analysis) and classification tasks (random forests and logistic regression), with the intention of identifying similarities between the alternating verbs as well as assessing possible predictors of object null instantiation. Overall, this paper sheds light on the psychological basis of omission, its usage conditions and the possibility of contact-induced language change in L2-Englishes.

References

Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford Linguistics. Oxford: Oxford University Press.

Diessel, Holger. 2019. *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*. Cambridge: Cambridge University Press.

Diessel, Holger. 2020. A Dynamic Network Approach to the Study of Syntax. *Frontiers in Psychology* 11: 1–15.

Fillmore, Charles J. 2007. Valency Issues in FrameNet. In *Valency: Theoretical, Descriptive and Cognitive Issues*, edited by Thomas Herbst and Karin Götz-Votteler, 129–60. Berlin; New York: De Gruyter.
Fillmore, Charles J. and Paul Kay. 1995. *Construction Grammar*. Stanford: CSLI Publica- tions.
Goldberg, Adele E. 2001. Patient Arguments of Causative Verbs Can Be Omitted: The Role of Information Structure in Argument Distribution. *Language Sciences* 23 (4): 503–24.
Herbst, Thomas, David Heath, Ian F. Roe and Dieter Götz. 2004. *A Valency Dictionary of English: A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Berlin: Mouton de Gruyter.
Hoffmann, Thomas. 2022. *Construction Grammar: The Structure of English*. Cambridge: Cambridge University Press.
Lambrecht, Knud and Kevin Lemoine. 2005. Definite Null Objects in (Spoken) French: A Construction-Grammar Account. In *Grammatical Constructions: Back to the Roots*, edited by Mirjam Fried and Hans C. Boas, 13–55. Amsterdam: Benjamins.
Lemmens, Marten. 2006. More on Objectless Transitives and Ergativization Patterns in English. *Constructions* Special Volume 1.
Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago & London: The University of Chicago Press.
Lyngfelt, Ben. 2012. Re-Thinking FNI: On Null Instantiation and Control in Construction Grammar. *Constructions and Frames* 4 (1): 1–23.
Ruppenhofer, Josef. 2004. *The Interaction of Valence and Information Structure*. University of California, Berkeley: eScholarship.
Ruppenhofer, Josef and Laura A. Michaelis. 2010. A Constructional Account of Genre-Based Argument Omissions. *Constructions and Frames* 2 (2): 158–84.

Onomastic referencing strategies in a corpus of seventeenth-century grammars of English

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

The 17th century represents an eventful time with regards to language development and instruction. The English language expanded to areas of language use where the classical languages had previously predominated in the late 17th century, which led to increased standardization in language usage (Nevalainen 2006, 42). This and other sociopolitical developments sparked a shift in favor of English being recognized as a separate academic discipline (Beal 2004, 102). The examination of onomastic references, i.e. name-based references, in the grammatical literature of the 17th century offers a unique perspective on language usage, sociolinguistics, and the cultural nuances embedded in linguistic artifacts of this period.

Such examinations of English grammar writing lie at the core of the HeidelGram project (http://heidelgram.de). Previous studies within the project have investigated linguistic means employed by 16th- and 19th-century grammarians when referring to other persons within their works, ultimately aiming for a full diachronic perspective. Most recently, the types of persons referenced by grammarians of the 17th century have been investigated. The present study further quantitatively and qualitatively analyzes the types of references made by 17th-century grammar authors. This allows us to identify where the authors position themselves in relation to others as well as changing or stable trends in referencing strategies.

The HeidelGram corpus, carefully curated to encompass a representative selection of grammatical works from the 16th to 19th centuries, serves as a valuable source for understanding how name-based references were employed in linguistic instruction at the time. For this study,

onomastic references within the 17th-century component of the HeidelGram corpus were systematically extracted and visualized in a citation network (see White 2012) using a custom-built tool based on Python and R. The 17th-century component of the corpus encompasses 17 texts, which add up to about 590.000 tokens. From these texts, a total of 2586 onomastic references were extracted. Each reference to a person was manually analyzed and assigned a reference category. There are six reference categories, which were originally established for the 19th-century grammar data (see Busse et al. 2018, 2019, 2020), such as *opinion* and *quotation*. The applicability of this categorization to the 17th-century data will be evaluated via inter-rater reliability measures.

Our previous work on the 16th- and 19th-century grammar books (Busse et al. 2018, 2019, 2020) has portrayed the potential of utilizing network analysis – a methodological tool for mapping relationships and patterns, as shown in the pilot network text analysis study on a sample of 17th century letters compiled from the Early Modern Letters Online (EMLO) (McGillivray and Sangati 2018). The application of network analysis allows us to construct and visualize the intricate connections between onomastic references within the grammar books. By mapping these linguistic networks, we aim to uncover patterns, clusters, and semantic relationships that contribute to a deeper understanding of the language norms and perceptions in 17th-century English. Our predictions are that the 17th-century component of the corpus shall include even more direct quotations than the 16th century due to the increased availability of printed books where authors could directly cite another person's work.

The reference strategies employed by grammarians to reference other authors show us how they position themselves with regards to certain beliefs and paradigms. The study elucidates the sociolinguistic dynamics of the time by revealing patterns in the selection and representation of names within the grammatical discourse. The categorization of onomastic references allows for an exploration of the social, cultural, and historical dimensions embedded in the linguistic fabric of the 17th century.

References

Beal, Joan C. 2004. *English in Modern Times.* London: Arnold.

Busse, Beatrix, Ingo Kleiber, Nina Dumrukcic and Sophie Du Bois. 2021. A corpus-based network analysis of 16th-century British grammar writing. CL2021, Limerick, Ireland, 2021.

Busse, Beatrix, Kirsten Gather and Ingo Kleiber. 2018. Assessing the Connections Between English Grammarians of the Nineteenth Century: A Corpus-Based Network Analysis. In *Grammar and Corpora 2016*, edited by Eric Fuß, Marek Konopka, Beata Trawiński and Ulrich H. Waßner, 435–42. Heidelberg: Heidelberg University Publishing.

Busse, Beatrix, Kirsten Gather and Ingo Kleiber. 2019. Paradigm Shifts in 19th-Century British Grammar Writing: A Network of Texts and Authors. In *Norms and Conventions in the History of English*, edited by Birte Bös and Claudia Claridge, 49-72. Amsterdam: John Benjamins.

Busse, Beatrix, Kirsten Gather and Ingo Kleiber. 2020. A Corpus-Based Analysis of Grammarians' References in 19th-Century British Grammars. In *Variation in Time and Space: Observing the World Through Corpora*, edited by Anna Cermakova and Markéta Malá, 133-172. Diskursmuster - Discourse Patterns 20. Berlin: De Gruyter.

McGillivray, Barbara and Federico Sangati (2018). Pilot study for the COST Action "Reassembling the Republic of Letters": Language-driven network analysis of letters from the Hartlib's Papers.

Nevalainen, Terttu. 2006. *An Introduction to Early Modern English*. Edinburgh: Edinburgh University Press.

'They cost me less pains than tragedy does': On the configuration of the quantifiers *less* and *fewer* in the history of English

Javier Calle-Martín and Marta Pacheco-Franco (University of Málaga)

Even though *less* and *fewer* are supposed to have their particular domains for the expression of quantification, this has not been always the case in the history of English. The use of *less* with countable nouns dates back to the Old English period (OED, s.v. *less*, adj. and adv.) and it was not until the late fourteenth century when the comparative form of *few* is first recorded in writing (OED, s.v. *fewer*, adj., P.2.a.). These two quantifiers have coexisted in complementary distribution since then, irrespective of the typology of the noun, up to the sixteenth century, although the alternation becomes more frequent with count nouns (*less things* vs *fewer things*) than with mass nouns *(less strength* vs *fewer strength*). The advent of prescriptivism brought about some sort of fresh air to the distribution of these quantifiers and it was Robert Baker (1770) who first postulated that *fewer* (and not *less*) should be exclusively used with count nouns. This prerogative was taken as a desideratum for the subsequent generations of grammarians to the extent that it has been dogmatic since then.

The issue stands out as controversial with a number of moot points still unanswered. On the one hand, on historical grounds, it is elsewhere noted that the eventual configuration of *fewer* and *less* is a typical change from above derived from the grammarians' prescriptive bias (Peters 2004: 205). This, however, raises questions as to whether the distinction between these quantifiers might have already been on the rise before the publication of Baker's work, thus opening the door to the possibility that the standardisation of *fewer* and *less* was a usage-derived development later confirmed by the grammarians' attentive glance. More important, on the other hand, is the uncertain state of the art in present-day usage, which "is not as strict at all [since] the rules of the mavens are made-up and idiosyncratic" (van Gelderen 2006: 229). Indeed, while the rule for *fewer* stands, *less* is used more fluidly regardless of the typology of the noun (Merriam-Webster 1995: 592), which seems to answer to structural pressures in analogy with the use of *more* (Denison 1998: 124; Denison and Hogg 2006: 38). In spite of the widespread acceptance of these claims, there is, to our knowledge, no empirical evidence supporting them.

In light of this, the present paper first investigates the distribution of these quantifiers in the period 1650-1800, to shed light on their historical development and to ascertain – by means of a precept-corpus analysis – the role possibly played by grammarians in the configuration of the system. Secondly, the study addresses the distribution of these items in Present-day English to evaluate the use of *less* in combination with count nouns in some varieties of English worldwide. The results tentatively point to a recurrent oscillation of *less* with count and mass nouns both in the eighteenth and twenty-first centuries, a finding which may corroborate that, if there was ever an established norm, be it from above or from below, it was very short-lived. As a corpus-based study, the source material comes from the *Early English Books Online* (*EEBO*) and *Eighteenth Century Collections Online* (*ECCO*) corpora for the historical period 1650-1800 and from the British English component of the *Global Web-based English* (*GloWbE*) corpus for the assessment of present-day usage.

References

Baker, Robert. 1770. *Reflections on the English Language*. London: J. Bell.

Denison, David. 1998. Syntax. In Suzanne Romaine (ed.), *The Cambridge History of the English Language. Volume IV: 1776-1997*. Cambridge: Cambridge University Press, 92–329.

Denison, David and Richard Hogg. 2006. Overview. In Richard Hogg and David Denison (eds.), *A History of the English Language*. Cambridge: Cambridge University Press, 1–42.

van Gelderen, Elly. 2006. *A History of the English Language*. Amsterdam/Philadelphia: John Benjamins.

Merriam-Webster. 1995. *Merriam-Webster's Dictionary of English Usage* (2nd ed., p. 592).

OED = *Oxford English Dictionary*. Oxford: Oxford University Press. http://www.oed.com/.
Peters, Pam. 2004. *The Cambridge Guide to English Usage.* Cambridge: Cambridge University Press.

A diachronic register-approach to complex prenominal modifiers

Marcus Callies[1], Turo Vartiainen[2] and Aatu Liimatta[2]

([1]University of Bremen, [2]University of Helsinki)

Prenominal modifiers have become increasingly common in the recent history of both British and American English (e.g. Biber et al. 2009; Biber & Gray 2016; Leech et al. 2009). This growing preference for premodification has been discussed in light of an ongoing trend towards a more compressed style of writing in some written registers of English, such as academic writing and journalistic prose, which has been interpreted to arise from a need to condense information in an economical way (sometimes referred to as "densification", see Leech et al. 2009: 234, 249). More specifically, in a recent corpus-based study on American English, Günther (2018) observed a notable frequency increase in highly complex premodifiers instantiating the A-to-V construction with *tough*-predicates (1) and the comparative *than*-construction (2).

(1) Below is my very simplified and **easy-to-understand** guide […]. (GloWbE, KE)
(2) What's this sleek, sexy Tokyo surprise doing in the **less-than-trendy** area of Sa Ying Pun? (GloWbE, HK)

In this paper, we examine the diachronic development and global spread of these two constructions in light of the hypothesis that complex prenominal modifier constructions are a register feature of magazines and news reporting, two registers that are driven by the need for densification. According to Biber and Conrad (2019), register features are words or grammatical characteristics that are: (1) pervasive, i.e. they are distributed throughout a text from the respective register; (2) frequent, i.e. they occur more commonly in the target register than in most comparison registers (but are not restricted to the target register); and (3) they are functionally motivated in that they serve important communicative functions in the target register. To examine the long-term and short-term diachronic trends of the A-to-V and the comparative *than*-constructions, we investigated data from two large, genre-balanced corpora of American English: the *Corpus of Historical American English* (COHA) and the *Corpus of Contemporary American English* (COCA). Our findings suggest that the rise of the two constructions in American English (AmE) in the second part of the 20th century (COHA data) is driven by the registers of MAGAZINES and NEWS. While frequencies appear to be tailing off in the past 30 years (COCA data), the two constructions are nevertheless much more frequently used in MAGAZINES and NEWS when compared to the other registers represented in the COCA. We can thus conclude that prenominal modifiers can indeed be considered register features of magazines and news reporting in American English.

To complement our AmE data, we examined the *News on the Web* corpus (NOW; web-based newspapers and magazines) in order to find evidence for the global spread of the two constructions in a comparable online register beyond American English (a topic that has received little attention in previous research on complex prenominal modifiers; however, see Mazaud 2004). Our results show that the constructions are indeed used globally in online newspapers and magazines, but they are much more frequent in South East Asian varieties of English when compared to African Englishes, a tendency that is at least partially affected by the typological profiles of the main L1 substrate languages.

References
Biber, D., J. Grieve & G. Iberri-Shea. 2009. Noun phrase modification. In G. Rohdenburg & J. Schlüter (eds.), *One Language, Two Grammars? Differences between British and American English*, 182–193. Cambridge: CUP.
Biber, D. & S. Conrad. 2019. *Register, Genre, and Style*. 2nd edition. Cambridge: CUP.
Biber, D. & B. Gray. 2016. *Grammatical Complexity in Academic English*. Cambridge: CUP.
Günther, C. 2018. A difficult to explain phenomenon: Increasing complexity in the prenominal position. *English Language and Linguistics* 23: 645–670.
Leech, G., M. Hundt, C. Mair & N. Smith. 2009. *Change in Contemporary English: A grammatical study*. Cambridge: CUP.
Mazaud, C. 2004. *Complex Premodifiers in Present-Day English. A corpus-based study*. PhD dissertation, University of Heidelberg, Germany.

Clause types and discourse modes: *There*-clauses in Late Modern historiography

Claudia Claridge (University of Augsburg)

There-clauses are a marked clause type introducing new information into the discourse (cf. Birner & Ward 1998). They have so far been studied in narrative (e.g. Adam 2012) and in academic (e.g. Jiang & Hyland 2020) contexts; in the latter context they are more frequently used in the humanities. Within this domain, historiography presents a context in which discourse modes (Smith 2003) such as narrative, description and argument intertwine and which probably shows a development over time from a more to a less narrative type (Claridge forthc.). The new information introduced by *there*-clauses in historiography can concern historical events as in (1)-(2), illustrating the existential and presentational type respectively, and form part of the historical narrative.

(1) But he (…) went on Conquering more Places in Scotland during the greatest Part of the ensuing Summer, when there was another Truce concluded between him and King Edward for some Months. (Tyrell, 1704)
(2) There remained, after the truce, no business of importance to detain Richard in Palestine; (Hume, 1770)

Alternatively, the new information may be part of the discussion of the author, involving explanations, arguments etc., cf. (3).

(3) It appears, consequently, that there are two ways of remedying excessive competition; either by increasing the whole annual produce of the country, or diminishing the number of competitors in all classes. (Wakefield, 1849)

The data for this study is taken from the *Corpus of Late Modern English Historiography*, containing writings by 50 historians from the period 1700 to 1914, and amounting to about 1.5 million words. Its overarching question concerns whether or how *there*-clauses and the text-typological development of historiography are correlated, i.e. that the occurrence of a certain type of *there*-clause may be taken as an indicator of the prevailing discourse mode. The research questions and some preliminary findings are:

1) How frequent are *there*-clauses generally and over time in historiography? How common are formal realisations, such as existential type, present-tense or modal verb (potentially more descriptive, argumentative), or presentational, past-tense or perfect verb (potentially more narrative)? The overall frequency (123 per 100,000) is remarkably close to that for Applied Linguistics (Jiang & Hyland 2020), while a clear temporal development is not visible, but instead drastic differences between individual historians, which indicates different discourse-mode and discipline-internal orientations. Tense choices show an almost equal split across the whole corpus and the presentational type is almost as common as in Adam's (2012) fictional data (6.5% vs. 7.5%). Type of noun phrase and presence of adverbials will also be investigated, with cases such as (1)-(2) again pointing to more narrative contexts.

2) What are there the local discourse functions in history? Do these change over time? Jiang & Hyland's most frequent function of asserting (non)existence may be the most frequent here too, cf. (1). Their rare function of "mark enumerations" is illustrated by (3), while their summarizing function has not been identified. The fairly high occurrence of negated types (c. 15%) warrants a closer look, as this may lead to a stance-related function.

References

Adam, Martin. 2012. Existential *there*-construction as a means of presentation in narrative (a corpus-based syntactic-semantic analysis). *Linguistica Pragensia* XXII (1): 1-17.

Birner, Betty J. and Gregory Ward. 1998. *Information Status and Noncanonical Word Order in English*. Amsterdam: Benjamins.

Claridge, Claudia. Forthc. History writing. In Merja Kytö and Erik Smitterberg (eds.), *New Cambridge History of the English Language.* Volume II: *Documentation, sources of data and modelling*. Cambridge: CUP.

Corpus of Late Modern English Historiography. 2019. Compiled by Sebastian Wagner. University of Augsburg. (In-house corpus)

Jiang, Feng (Kevin) & Ken Hyland. 2020. "There are significant differences…": the secret life of existential there in academic writing. *Lingua* 233 (Jan.) (no pagination).

Smith, Carlotta. 2003. *Modes of Discourse. The Local Structure of Texts.* Cambridge: CUP.

New kids on the (corpus) block: Introducing TaCoCASE and WebCorpLSE

Caroline Collet[1], Stefan Diemer[1], Matt Gee[2] and Andrew Kehoe[2]

([1]Umwelt-Campus Wirkenfeld, [2]Birmingham City University)

The aim of this paper is threefold: (1) it introduces a new spoken CMC corpus TaCoCASE, and (2) a new search tool WebCorpLSE, in which the corpus is integrated, and (3) it demonstrates functionality and applications through a case study.

TaCoCASE (Transatlantic Component of the Corpus of Academic Spoken English) is a corpus compiled as part of the CASE project – released in September 2023. The corpus consists of 15 computer-mediated conversations (CMC) between native and non-native English speaking students from the UK, Germany and the United States. The total length of the conversations is 10.5 hours or 140,003 tokens. TaCoCASE can be used in combination with ViMELF, another sub-corpus from the CASE project, as it adds a native speaker component to the data. Due to its multimodal set-up as a spoken CMC corpus, TaCoCASE is a rich data source which facilitates research in many different fields.

Both TaCoCASE and ViMELF have been integrated into a new version of the WebCorp Linguist's Search Engine (WebCorpLSE), which is being presented here for the first time. WebCorpLSE provides access to corpora built by crawling the web and now acts as an online search and analysis tool for other corpora too. All corpora are annotated using the Stanford CoreNLP tools (Manning et al. 2014), and include lemma annotations and part-of-speech categories based on the Universal Dependencies framework.

This paper focuses on TaCoCASE and we begin by exploring frequently occurring lemmas in the corpus. One example is *REALLY*, which appears in all 15 conversations with an overall frequency of 2,150 per million words. We use WebCorpLSE to examine its collocates, most of which are adjectives preceded by *REALLY* as an intensifier: *NICE*, *BIG*, *COOL*, *GREAT*, *BAD*, *HARD*, *FUN*. Interestingly, *REALLY* collocates with itself 82 times, reflecting the frequent use in TaCoCASE of repetition for emphasis, e.g. "a family friend of ours uhm was really suffering from really really bad depression".

A new WebCorpLSE feature is a user-friendly display of differences between the collocational profiles of word pairs. This is of particular benefit in language teaching when exploring usage variation between semantically-related words, drawing upon research on lexical repulsion (Renouf & Banerjee 2007). We illustrate this feature by presenting, amongst other examples, the profiles of the adverbs *QUITE* and *PRETTY*, which demonstrate both overlapping collocates (*GOOD*, *COOL*) and unique collocates (*QUITE* with *INTERESTING* and *HARD*; *PRETTY* with *COMMON*).

Following a corpus-based discourse analytical approach we conduct a qualitative analysis by exploring strategies the participants use to handle intercultural communication and prevent misunderstandings. Examples from the corpus show how the participants mediate these intercultural situations when explaining new concepts. In terms of communicative strategies, participants frequently use code-switches in order to explain a concept from their own culture. They also break down complicated information, paraphrase, spell out and link to previous knowledge when explaining concepts new to their interlocutor. These examples are ideal for enhancing intercultural competence in teaching or business contexts.

References

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Renouf, Antoinette and Jayeeta Banerjee. 2007. The search for repulsion: a new corpus analytical approach. In *VARIENG - Studies in Variation, Contacts and Change in English* 2*: Towards Multimedia in Corpus Studies*. University of Helsinki.

TaCoCASE. 2023. Transatlantic Component of the CASE project. Birkenfeld: Trier University of Applied Sciences. Version 1.0. Compiler: Collet, Caroline. [http://umwelt-campus.de/case/TaCoCASE] (14 November 2023)

ViMELF. 2018. Corpus of Video-Mediated English as a Lingua Franca Conversations. Birkenfeld: Trier University of Applied Sciences. Version 1.0. The CASE project [umwelt-campus.de/case]. (14 November 2023).

Predicting the CEFR level of English listening texts with machine learning methods

Christopher Cooper (Rikkyo University)

The comprehension of listening texts tends to be judged by lexical coverage, based on figures such as learners need to know 95% (van Zeeland & Schmitt, 2013) or 90% (Durbahn et al., 2020) of the words in a text to understand it. However, this is not very easy to interpret for language teachers. The Common European Framework of Reference (CEFR) is increasingly influential in the field of language learning, teaching and assessment across Europe, and it is beginning to have an influence in Asian countries such as Japan and Vietnam (Tono, 2019). As learners are often put into classes based on proficiency level, a CEFR level is likely to be more interpretable when judging the difficulty of listening texts. Tools have been created to estimate the CEFR level of reading texts beyond lexical coverage in Chinese (Sung et al., 2015) and English (Uchida & Negishi, 2018). In addition, Machine learning methods have been used to predict the CEFR level of English learner writing using the large language model BERT (Schmalz & Brutti, 2022). So far, no such studies have been conducted for listening texts. The current study hopes to bridge this gap by investigating the potential to predict the CEFR level of listening texts. This is part of a larger project in which one of the goals is to predict the CEFR level of YouTube videos.

Research questions:

1) How accurately can machine learning methods predict the CEFR level of listening texts?
2) Which method has the highest accuracy?

To answer the research questions, a corpus of CEFR-labelled listening texts was created. 563 texts (around 300,000 words) were scraped from the British Council website (https://learnenglish.britishcouncil.org/) and 153 texts (around 40,000 words) were included from sample tests of the Cambridge English exams (https://www.cambridgeenglish.org/learning-english/exam-preparation/). The CEFR labels were assigned by the materials creators, both well-known material developers in the ELT industry. Although the corpus contains texts that were created for language learning, 247 of the British Council texts were YouTube videos, and a further 143 texts were language learning videos, which goes some way to fulfilling the eventual goal of the corpus, to predict the level of YouTube videos. The corpus size was small compared with previous projects due to the availability of CEFR-labelled listening texts.

As this was an exploratory project, three types of variables were created from the corpus data to see which method was the most accurate. The first method used information about the complexity of the grammar and vocabulary in the text, and the speed of speech; variables that have been shown to affect listening comprehension (Bloomfield et al., 2010). The other methods used text embeddings, which represent the semantic meaning of the texts. Specifically, I used a BERT transformers model and Chat GPT embeddings. Four different machine learning methods were used with Scikit Learn (Pedregosa et al., 2011) in Python. The specific methods were Random Forests, Support Vector Machines, K Nearest Neighbours, and a neural network MLP Classifier.

Initially the data were split into 5 classes, A1, A2, B1, B2, and C level. As none of the methods could accurately distinguish between B2 and C level, the approach was adapted to only include four classes: A1, A2, B1, and B2+. Various parameters were trialled, including the bootstrapping of texts in the training dataset. The accuracy of each method was evaluated by comparing the predicted label in the test data with the actual label from the original text. The most accurate method used Chat GPT embeddings, Support Vector Machines and bootstrapping. The overall

accuracy was 0.81, with macro averages of precision = 0.75, recall = 0.78, and f-score = 0.76. This method has the potential to predict the CEFR level of listening texts. However, the accuracy could be improved, and future research could use larger datasets.

References

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J. & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension*: Defense Technical Information Center. https://doi.org/10.21236/ADA550176

Durbahn, M., Rodgers, M. & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System* 88, 102166. https://doi.org/10.1016/j.system.2019.102166

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12(85), 2825–2830.

Schmalz, V. J. & Brutti, A. (2022). Automatic assessment of English CEFR levels using BERT embeddings. In E. Fersini, M. Passarotti & V. Patti (Eds.), *Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-it 2021* (pp. 293–299). Accademia University Press. https://doi.org/10.4000/books.aaccademia.10828

Sung, Y., Lin, W., Dyson, S. B., Chang, K. & Chen, Y. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal* 99(2), 371–391. https://doi.org/10.1111/modl.12213

Tono, Y. (2019). Coming full circle: From CEFR to CEFR-J and back. *CEFR Journal* 1, 5–17.

Uchida, S. & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. In Y. Tono & H. Isahara (Eds.), *Proceedings of the 4th Asia Pacific corpus linguistics conference (APCLC 2018)* (pp. 463–468).

Van Zeeland, H. & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics* 34(4), 457–479. https://doi.org/10.1093/applin/ams074

Communicating science to non-specialist audiences: A multidimensional analysis of scientific blogs

Niall Curry and Pascual Pérez-Paredes

(Manchester Metropolitan University, University of Murcia)

Parascientific discourses include a wide-range of registers, spanning scientific blogs, research project websites, specialist magazines, and podcasts (Pérez-Llantada, 2021). In recent years, there has been a growing interest in research on parascientific communication (e.g., Curry & Pérez-Paredes, 2021; Pérez-Llantada, 2021; Zou & Hyland, 2019), largely owing to the increased value placed on communicating science to the wider public (Liao et al., 2020). Scientific blogs are gaining popularity as they can serve as a means to make research accessible to diverse audiences (Pérez-Llantada, 2021). Yet, given their relative novelty, scientific blogs remain a somewhat fuzzy register and there is a limited understanding of what makes for a typical scientific blog. Some studies on scientific blogs have focused on describing variation in their linguistic features across disciplines in terms of the use of engagement markers, for example (Zou & Hyland, 2020). In studying how knowledge of global issues, such as the COVID-19 pandemic, is socially and discursively constructed in scientific blogs, research has shown that writers communicate differently about topics of public interest when compared to specialised, esoteric subject matter (Curry & Pérez-Paredes, 2021). Recognising the value of scientific blogs for public engagement and our limited understanding of how such forms of communication are

socially constructed, a holistic view of the register of scientific blogs is needed. Such insight would allow us to understand how professional writers communicate complex research to non-specialist audiences and how this varies across disciplinary area and thematic focus.

Addressing this need, this paper presents a register analysis of scientific blogs published in The Conversation – an international website dedicated to publishing academic blogs on a range of subjects in a range of languages. The conversation is a valuable source for studying science communication as it is a medium through which researchers can engage with non-specialist audiences and one which undergoes an editorial process that ensures comparability of text in terms of publishable quality (Curry & Pérez-Paredes, 2021). Moreover, owing to the editorial process, blogs from The Conversation are less likely to exhibit the dialogicity found in personal blogs (Bondi, 2022). Undertaking a multidimensional analysis of scientific blogs (e.g., Biber & Egbert, 2016), this study examines a corpus of academic blogs related to the climate crisis with a view to determining the sub-registers that compose the fuzzy genre of academic blogs in The Conversation. While wholesale analysis exhibits much variability, through a thematic focus, the analysis identifies a number of dimensions that offer insight into the register features of academic blogs.

Overall, the contributions of the study are threefold. First, the findings tell us more about how scientific writers write blogs for non-specialist audiences. Second, the findings offer a point of reflection on the affordances of multidimensional analysis for the interrogation of one broadly conceived yet undefined register. Third, the findings offer guidance for communicating science to non-specialist audiences, with a specific focus on communicating global crises.

References

Bondi, M. (2022). Dialogicity in individual and institutional scientific blogs. *Publications* 10(1), 23-42.

Biber, D. & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics* 44(2), 95-137.

Curry, N. & Pérez-Paredes, P. (2021). Stance nouns in COVID-19 related blog posts: A contrastive analysis of blog posts published in The Conversation in Spain and the UK. *International Journal of Corpus Linguistics* 26(4), 469-497.

Liao, Q., Yuan, J., Dong, M., Yang, L., Fielding, R. & Lam, W. (2020). Public Engagement and government responsiveness in the communications about COVID-19 during the early epidemic stage in China: Infodemiology study on social media data. *Journal of Medical Internet Research* 22(5), 1-13.

Pérez-Llantada, C. (2021). *Research genres across languages: Multilingual communication online*. Cambridge University Press.

Zou, H. & Hyland, K. (2019). Reworking research: Interactions in academic articles and blogs. *Discourse studies* 21(6), 713-733.

Zou, H. J. & Hyland, K. (2020). "Think about how fascinating this is": Engagement in academic blogs across disciplines. *Journal of English for Academic Purposes* 43, 1-12.

Hella entrenched and *hella* (un)conventional: The cognitive-sociolinguistics of *hella*-intensification

Robert Daugs (University of Kiel)

The intensifier *hella* is rather (in)famous for its flexible syntactic distribution, its status as a regional shibboleth of Northern California and the Bay Area, and its etymology. Unsurprisingly, it has prompted some serious interest among syntacticians, (perceptual) dialectologists, and (diachronic) construction grammarians (see e.g. Adams 2009; Boboc 2016; Bucholtz 2006; Bucholtz et al. 2007; Hoffmann & Trousdale 2011; Russ 2013; Trousdale 2012; Wood 2019).

Two issues (at least) regarding *hella* are arguably unresolved. First, from a cognitive perspective, the emergence of *hella* appears to be a straightforward case of coalescence brought about by an increase in usage intensity of the presumably original utterance *hell of a*, which eventually led to automatization and phonetic reduction: *hell of a* > *helluva* > *hella*. This fails to explain, however, why *hella* and its source form exhibit noteworthy frequency asymmetries in their complementation patterns. Unlike *hell of a* and *helluva*, which prototypically modify nouns, *hella* seems to mainly modify adjectives (e.g. *hella good*, *hella righteous*; here competing with more conventional intensifiers like *very* or *really*) and only more recently nouns (e.g. *hella things*, *hella people*; here competing with *a lot of* or *many*) in PDE. Also, historically, the most frequent element to follow both *hell of a* and *helluva* is *lot*, while *hella lot* is comparatively rare. How do we get from *(a) hell of a* N to *hella* ADJ?

Second, from a social perspective, while *hell of a* is fully conventionalized in American English in general and should thus have roughly the same licensing potential across different dialects, the fully reduced form *hella* remains regionally stratified, its pop-culture induced awareness notwithstanding (cf. Daugs 2019). This suggests that conformity to social order on the one hand and entrenched pragmatic associations on the other play a crucial role in boosting or inhibiting the usualization of *hella*. To elaborate, speakers from a *hella*-community utilize the form to establish and maintain their social identity. The more often they do so in usage events, the more strongly the association between the form and the communicative goal of signaling belonging ('that is how WE speak') becomes entrenched. Conversely, speakers from an 'outside' community, while being aware of *hella* via diffusion, consciously avoid its use ('that is how THEY speak'), thereby prohibiting the form from becoming established. On the cognitive side, this avoidance must also be stored in terms of entrenched pragmatic associations, yet not by means of frequent repetition.

The goal of the present study is to account for both issues by investigating the interaction between the social and cognitive dimension of *hella*-intensification. Based on historical/contemporary corpus and web-based data of American English, we will revisit the emergence and conventionality of *hella* relative to its parent forms as well as its phonetic and functional relatives *sorta*, *kinda*, and *lotta*, on whose routinization and diffusion the intensifier may have very well piggybacked. On the theoretical end, the study draws on Schmid's (2020) *Entrenchment-and-Conventionalization Model*, which unifies the individual and communal level of language.

References

Adams, Michael. 2009. *Slang: The people's poetry.* Oxford: Oxford University Press.

Boboc, Wellesley. 2016. *To hella and back: A syntactic analysis of hella in dialects of American English*. New York: New York University Senior Honors Thesis.

Bucholtz, Mary. 2006. Word up: Social meanings of slang in California youth culture. In Jane Goodman and Leila Monaghan (eds.), *A cultural approach to interpersonal communication: Essential readings*, 243–267. Chichester: Wiley-Blackwell.

Bucholtz, Mary, Nancy Bermudez, Victor Fung, Lisa Edwards & Rosalva Vargas. 2007. *Hella Nor Cal or totally So Cal?* The perceptual dialectology of California. *Journal of English Linguistics* 35(4), 325–352.

Daugs, Robert. 2019. *Get your ICAME on*: Constraints, expansion and productivity of GET POSS X *on*. Paper presented at ICAME40, June 1–5, Neuchâtel.

Hoffmann, Thomas & Graeme Trousdale. 2011. Variation, change and constructions in English. *Cognitive Linguistics* 22(1), 1–23.

Russ, Robert Brice. 2013. *Examining regional variation through online geotagged corpora*. Columbus: The Ohio State University Master's thesis.

Schmid, Hans-Jörg. 2020. *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment.* Oxford: Oxford University Press.

Trousdale, Graeme. 2012. Grammaticalization, constructions and the grammaticalization of constructions. In Kristin Davidse, Tine Breban, Lieselotte Brems and Tanja Mortelmans (eds.), *Grammaticalization and Language Change: New Reflections*, 167–198. Amsterdam: John Benjamins.

Wood, Jim. 2019. Quantifying geographical variation in acceptability judgments in regional American English dialect syntax. *Linguistics* 57(6), 1367–1402.

Reconstructing American English inputs in a globally available mass media product: Intensifiers in the television series *Gilmore Girls*

Julia Davydova (University College of Teacher Education Vorarlberg)

The role played by mass media in the propagation of the patterns of structured variability has developed into one of the most challenging and hotly debated issues in modern sociolinguistics (Stuart-Smith 2007; Bell & Sharma 2014). Current epistemological models (Sayers 2014) propose that the study of possible media effects on (non-)acquired speech patterns must involve systematic comparisons across source speech communities, adopting speech communities and mass media texts. Against this backdrop, the study sets out to explore language-specific and sociolinguistic conditioning underlying the use of intensifiers in the television series *Gilmore Girls* and compares it to that of L1 English and English spoken as a foreign language (EFL). Firmly grounded in the methodological paradigm of variationist sociolinguistics, this investigation pinpoints some indisputable similarities in the use of intensifiers by fictional characters and real speakers of L1 English and EFL. These are revealed by the overall rates of intensification, the general make-up of the (most frequent) linguistic variants and their sociolinguistic conditioning. The system of language-internal conditioning triggering the realization of individual intensifiers is found to be quite distinctive from that reported for both L1 English vernaculars and EFL English. I discuss the possible contribution that these findings make to the existing models of L2 acquisition and language change, while also proposing directions for future research.

References

Bell, Alan & Devyani Sharma. 2014. Debate. Media and language change. *Journal of Sociolinguistics* 18(2): 213.

Sayers, Dave. 2014. The mediated innovation model: A framework for researching media influence in language change. *Journal of Sociolinguistics* 18(2): 185-212.

Stuart-Smith, Jane. 2007. The influence of the media. In Carmen Llamas, Louise Mullany and Peter Stockwell (eds.) *The Routledge Companion to Sociolinguistics*. London & New York: Routledge.

Beyond the transactional: Identifying and analysing expressive speech acts in workplace emails

Rachele De Felice (The Open University)

The main function of workplace emails is generally recognised to be transactional or task-oriented: asking for or providing information, making requests, agreeing to actions. However, these messages often also contain phatic utterances, like generic good wishes ('I hope you had a good Christmas') and more personal comments ('Sorry to hear you had a bad trip' or 'I'm feeling demoralised about this result'). This presentation discusses the range of expressive utterances found in two large workplace email datasets (the Enron Corpus and the Clinton Email Corpus), looking at the diversity of functions they embody and their potential effects on the

interpersonal relationships between their senders and recipients. Using Searle's definition of expressives – "Speech acts that express the speaker's feelings about themselves or the world" (Searle 1976: 12) – it asks:

1) How easy is it to identify expressives in workplace emails?
2) What are their functions within this register?
3) What can they tell us about workplace relationships?

Manual annotation of a subset of the Enron Corpus yielded over 1000 instances of expressives, covering a range of both positive and negative emotions, e.g.:

(1) **Very glad** to hear that things have gone well so far.
(2) **I'll keep my fingers crossed**.
(3) **Really sorry** to bother you so much.
(4) **We respect** the effort.

These then underwent manual functional annotation, using a scheme based on previous work (Guiraud et al. 2011, Ronan 2015). The scheme uses eight categories representing emotions and functions: agreement, disagreement, thanking, apologising, sorrow, joy, greetings/good wishes, and intentions. Through this annotation, we find that the most frequent functions are highly routinised, superficial expressions of emotions such as greetings and good wishes (*thanks*, *I look forward to our visit*, *enjoy the weekend*, and so on), followed by utterances where the expressive phrase is mainly acting as a frame or downtowner for a different type of speech act (e.g. *I'm not convinced he was referring to Form AB1X*). Utterances conveying 'real' emotions such as happiness, displeasure, worry, are much less frequent, with negative feelings being particularly rare. Generally, more expressive adjectives are used to describe third parties rather than the speakers' own emotions.

The categories highlighted above form the basis of a qualitative analysis of emails from the Clinton Email Corpus, where the hierarchical and interpersonal relationships between interlocutors are known, to better understand how the different functions feature in different exchanges. Among salient findings, we observe that utterances with more genuine emotional content are a marker of a close relationship – not just of a reciprocal close relationship, but also of an aspiration to one. For example, individuals writing to Clinton make use of many expressives, but her replies rarely contain the same amount (if any). However, with members of her inner circle, Clinton freely expresses negative emotions such as disappointment and worry – thus also giving an insight what things worry her (weather, flight delays, schedules, email troubles…).

This research shows that, beyond the specific illocutionary act carried by an expressive, this speech act plays an important role in shaping and reflecting individuals' identities at work through communicative practices.

References

Guiraud, N., Longin, D., Lorini, E., Pesty, S. & Rivière, J. (2011). The face of emotions: a logical formalization of expressive speech acts. In *10th International Conference on Autonomous Agents and Multiagent Systems* (AAMAS 2011), 1031-1038.

Ronan, P. (2015). Categorizing expressive speech acts in the pragmatically annotated SPICE Ireland corpus. *ICAME Journal* 39(1), 25-45

Searle, J. R. (1976). A classification of illocutionary acts1. *Language in society* 5(1), 1-23.

A contrastive corpus study of the academic registers of primary and secondary school

Alice Deignan and Elena Semino (University of Leeds, Lancaster University)

The language of school differs from the everyday language that children encounter in their non-school lives (Gee, 2008), and that difference increases sharply at the point when they leave primary, or elementary school to start secondary, or high school. Primary school (ages 4-11 in England) typically follows a thematically-organised, child-centred curriculum, while secondary school (ages 11-16 or 18 in England) has a disciplinary-organised curriculum designed to prepare students for national high-stakes examinations. As well as increased academic demands, children encounter increasingly specialised language at this point. This can present a particular barrier for children from lower socio-economic status backgrounds, who are statistically more likely to falter academically as they start secondary school.

There are several detailed inventories of the language of schooling based on small-scale text analysis (e.g. Schleppegrell, 2001; Snow & Uccelli, 2009), but to date there had been no large-scale corpus analysis, a gap which we address. The research questions addressed in this paper are:

1) How does the lexis encountered in secondary school differ from that of primary school?
2) How does the lexis encountered in secondary school differ from the lexis children are likely to have encountered outside school?

To tackle these questions, we compiled corpora of written texts and teacher talk from (1) the last two years of primary school and (2) the first two years of secondary school, totalling 3.5 million words, from 13 primary and secondary schools in England. We also use the BNC2014 and BNC2014 (Spoken) as reference corpora. Using #LancsBox6 and Sketch Engine, we conducted a series of Key Word analyses comparing our secondary school and primary school corpora, both as whole corpora and comparing subcorpora of texts from the school subjects of mathematics, English and science. We also compared our secondary school corpus with reference corpora. The KW studies were supplemented with qualitative concordance analysis. Interviews with students provided context.

We found that while there are some completely new lexical items in secondary school, a large vocabulary learning load consists of new, metaphorical meanings of known words, and sometimes the reverse, less-known literal meanings of words which are more usually metaphors. They are often subject specific, and have specific patterns and forms. For example, in mathematics, children will encounter metaphorical uses such as *expand* [an equation], [square] *root*, *round* [number], as well as highly restricted uses of everyday words, such as problem. Science and mathematics vocabulary items such as *concentrate* (substance that is not watered down) and *prime* (a number that can only be divided by itself and one) are possibly literal counterparts to their more familiar everyday meanings 'think hard' and 'best quality'.

Our interview data suggest students have insufficient contextual information to work out the discipline-specific meanings of such words in enough detail to apply them academically. The research harnesses corpus and discourse tools to provide support with academic school language for students, teachers and materials developers.

References

Gee, J. P. (2008) What is academic language? In Roseberry, A., Warren, B. (eds) *Teaching Science to English Language Learners: Building on students' strengths*, NSTA Press, 57-70.

Schleppegrell, M. (2001) Linguistic features of the language of schooling. *Linguistics and Education* 12/4, 431-459.

Snow, C., Uccelli, P. (2009) The challenge of academic language. In Olson, D. R., Torrance, N. (eds.) *The Cambridge Handbook of Literacy*. Cambridge: Cambridge University Press, 112-133.

Examining the persuasive influence of implicitness and epistemic stance in newspaper and political discourse on immigration and humanitarian crises

Elena Domínguez-Romero[1], Marta Carretero[1] and Mercedes González-Vázquez[2]

([1]Complutense University of Madrid, [2]University of Vigo)

Implicitness, coupled with the epistemic stance, plays a pivotal role in persuasive communication, as elucidated by scholars such as Holtgraves (1998), Heritage and Raymond (2005), Lombardi and Masia (2014), and Marín-Arrese (2021). Building upon this idea, the present paper aims to investigate the relationship between implicit communication and impersonal expressions of epistemic stance.

Following Lombardi and Masia (2014: 161), we will distinguish between the implicitness of content (implicatures) and the implicitness of both content and responsibility (presuppositions). Both categories function to diminish the addressee's inclination toward critical reactions, as emphasized by Holtgraves (1998). Regarding implicatures, which are implicit meanings geared at the hearer's crafting of inferences, an expectation exists for belief in their truth. In contrast, presuppositions present the communicated content as knowledge shared and agreed upon by the addressee, while the speaker's responsibility remains implicit. Furthermore, epistemic stance relates to the "speaker/writer's endeavor to control conceptions of reality, involving their assessment of the truthfulness of the designated event and the likelihood of its occurrence, and/or their specification of the sources of information that entitle them to make a factual claim" (Marín-Arrese 2021: 290). Based on Marín-Arrese's (2021) approach, the present study comprises the following impersonal expressions of epistemic stance: (i) impersonal factives (*in truth / fact / reality, the truth / fact (is) that, it is known / remembered that*…), (ii) impersonal markers of cognitive attitude (*it is assumed / believed / thought that…, it is conceivable / plausible that*…), (iii) markers of interpretation of evidence (*it is shown that, the evidence / proof is that*…), and (iv) impersonal ignoratives (*it is not known that, nobody knows / understands that*…).

Our research question explores how impersonal epistemic expressions contribute to persuasiveness in two distinct registers: political discourse and opinion articles. We hypothesize that these expressions contribute to reducing critical reactions, akin to features shared with presuppositions. To decode the true intricacies inherent in the relationship between implicit communication and impersonal expressions of epistemic stance, we harness the analysis of two 60,000-word English corpora of conservative discourse addressing immigration and humanitarian crises. The corpora belong to two different registers: one contains newspaper opinion articles sourced from *The Telegraph*, and the other consists of spoken political discourse by members of right-wing parties, mainly the Conservative Party, UKIP and Reform UK. The analysis begins with a search for the aforementioned expressions of epistemic stance, combining manual and automatic procedures. It specifically targets the epistemic expressions that trigger presuppositions.

The findings indeed point towards the shared characteristics of impersonal expressions of epistemicity and implicit communication, which are discussed comparatively. This discussion emphasizes the role of impersonal epistemic expressions in enhancing persuasiveness, within newspaper and political discourse, while highlighting register disparities between the two corpora.

References
Boye, K. (2012). *Epistemic meaning. A crosslinguistic and functional-cognitive study*. Berlin / Boston: Mouton De Gruyter.
Holtgraves, T. (1998). Interpersonal foundations of conversational indirectness. In Fussell, S.R. & Kreuz, R.J. (eds.) *Social and cognitive approaches to interpersonal communication*. Mahwah, NJ.: LEA, 71-89.
Lombardi, E. & Masia, V. (2014). Implicitness impact: Measuring texts. *Journal of Pragmatics* 61, 161-184.
Marín-Arrese, J. I., (2013). Stancetaking and inter/subjectivity in the Iraq inquiry: Blair vs. Brown. In Marín-Arrese, J., Carretero, M., Arús, J., van der Auwera, J. (eds.) *English modality: Core, periphery and evidentiality*. Berlin: Mouton de Gruyter, 411-445.
Marín-Arrese, J.I. (2021). Winds of war: Epistemic and effective control in political discourse. *Cultura, lenguaje y representación* 26, 289-307.

How do probabilistic grammars develop in spoken EFL? The influence of proficiency level on the choice between *will* and *be going to*

Tanguy Dubois, Magali Paquot and Benedikt Szmrecsanyi (KU Leuven)

Previous research on alternation phenomena in English as a Foreign Language has mostly focused on how learners' mother tongue influences their choice of variant, while ignoring the impact of their proficiency level. Including proficiency level allows one to track the development of probabilistic grammars that guide the choice between variants at different stages of language learning. In this way, Dubois et al.'s (2023) study on the genitive alternation showed that possessor animacy, otherwise the strongest constraint in the genitive alternation, is weaker for low-proficiency learners of English, which they attribute to general learning mechanisms that apply regardless of the learners' mother tongue. In the present study, we investigate the impact of learners' proficiency level on the choice between *will* and *be going to* (*he will read the newspaper* vs. *he is going to read the newspaper*) in spoken language, which differs from the genitive alternation as it does not involve a change of word order. Additionally, investigating how proficiency level affects the choice of future marker forms a desideratum within research on the acquisition of the future markers, where learners' mother tongue does not appear to be highly influential (Bardovi-Harlig 2000: 411–412).

Methodologically, we collected 3616 instances of *will* and *be going to* from the Trinity Lancaster Corpus, a three-million-word corpus consisting of transcribed recordings from a spoken language exam between an examiner, who is a native speaker of British English, and low-intermediate to advanced learners of English from a wide variety of mother tongue backgrounds. The future marker observations were annotated for constraints known to probabilistically influence the choice of variant, including structural persistence, the type of sentence, clause, verb and subject, the presence of temporal adverbs, the temporal proximity of the future event and the length of the clause (see Engel & Szmrecsanyi 2023). The choice of variant was then analyzed using mixed-effects logistic regression, where the probabilistic constraints were entered as predictors in interaction with the speakers' proficiency level.

Results show that learners differ from native speakers regarding most relevant constraints at specific stages of language learning, regardless of their mother tongue background. Specifically, low-proficiency learners are sensitive to more constraints than native speakers, which is due to their more restricted usage of *be going to* for events that are relatively certain to happen in the near future. At the same time, these learners might be influenced by prescriptivist rules from English textbooks, which consistently cover the usage of the future marker variants (Burton

2023). By contrast, native speakers do not distinguish between the variants to the same extent, resulting in their more frequent use of *be going to*.

References

Bardovi-Harlig, Kathleen. 2000. Chapter Seven: Past, Present, and Future. *Language learning* 50(s1), 409–437.

Burton, Graham. 2023. *Grammar in ELT and ELT Materials: Evaluating its History and Current Practice* (Second Language Acquisition 164). Bristol, Blue Ridge Summit: Multilingual Matters.

Engel, Alexandra & Benedikt Szmrecsanyi. 2023. Variable grammars are variable across registers: Future temporal reference in English. *Language Variation and Change* 34(3), 355–378.

The use of the response form *uh-huh* in British conversation, TV transcripts and fictional dialogue: Affirmative answer, backchannel, or something else?

Jarle Ebeling (University of Oslo)

Work-In-Progress

This study investigates the use of the response form *uh-huh* in different registers. A response form is an insert defined by its pragmatic function, rather than by its inherent, semantic meaning (Biber et al. 1999: 1089*ff*). My interest in this particular response form is its growth in dialogue in fiction over the past century and the way it is deliberately used in the TV series "Succession" to signal an affirmative response with a "non-committal air" (Biber et al. 1999: 1091) to create a sense of uncertainty among the members of the Roy family.

According to the *Oxford English Dictionary*, *uh-huh* is "used to express assent or agreement, or as a non- committal response to a question or remark". Tolins & Fox Tree (2014) call *uh-huh* an acknowledgement token and a generic backchannel signalling that the speaker has the listener's attention and permission to go on with the narrative. In a recent article, Jucker (2021) looks at features of orality, including response forms, and shows how the frequency of these forms varies widely in the corpora he studies. They are predictably much more frequent in conversation than in fictional dialogue, with transcriptions of film and TV dialogue making up the space between the two extremes. Tottie (2017), investigating another feature of orality, i.e. the planners (hesitators) *uh*, *um* and *er* in American English, shows that these items carry different meanings and functions depending on the medium, written or spoken, and on their position within the clause. In written language, they are stance adverbs, expressing the writer's attitude, and in spoken interaction they function (mainly) as planners.

Inspired by Jucker and Tottie in particular, the study primarily addresses the following research question: 1) Can similar differences in use and frequency be detected for *uh-huh* as for the hesitators/planners? Moreover, since the material used for the study is drawn from the Spoken BNC2014 (Love et al. 2017), the Corpus of British Fiction (Ebeling, forthcoming), and the UK/IE part of the TV Corpus (Davies 2019), where we have access to the speakers' gender, a second research question will be: 2) To what extent does the gender of the speaker play a role regarding the use of this particular response form?

A preliminary look at the data indicates that there may indeed be differences between these three spoken registers in the use of the response form *uh-huh*. In the spoken BNC2014 it seems to be primarily used as a backchannel, while in dialogue in fiction it acts as an affirmative response to the preceding question or as a remark occurring in initial position in the clause or constituting a single-word sentence. Below are two examples.

(1) <u n="f1">and you feel comf you can go out mm for a pint during the day</u> <u n="m1">uh-huh</u> <u n="f1">and women can drink a whole a big pint big nice pint</u>
(2) “She seems nice, too.” “Uh-huh,” Lissa said, remembering how completely infatuated she'd once been with Antonio. (CBF)

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Davies, Mark. (2019) *The TV Corpus*. Available online at https://www.english- corpora.org/tv/.

Ebeling, Jarle. Forthcoming. 120 years of reporting clauses: stability or change? In Sarah Buschfeld, Theresa Neumaier, Patricia Ronan, Andreas Weilinghoff & Lisa Westermayer (eds), *Crossing Boundaries through Corpora: Innovative Approaches in Corpus-Linguistics*. John Benjamins.

Jucker, Andreas H. 2021. Features of orality in the language of fiction: A corpus-based investigation. *Language and literature* 30 (4), 341-360.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22 (3), 319-344.

Tolins, Jackson & Jean E. Fox Tree. 2014. Addresseee backchannels steer narrative development. *Journal of Pragmatics* 70, 152-164.

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

When, interestingly (enough), oddly *is enough*: A corpus-based study of *enough*-support in British and American English

Matthias Eitelmann (University of Mainz)

This paper focuses on the notion of *enough*-support, a grammatical variation phenomenon that concerns sentence adverbs such as *oddly* or *interestingly*. These may occur either on their own as in (1) and (3), or in combination with *enough* as in (2) and (4) (all examples taken from the US section of the GloWbE corpus).

(1) Oddly, as broad and general as the definitions of these crimes are, they would not reach cyberattacks.
(2) Oddly enough, my biggest obstacle right now is that I have a fairly secure career that I’m actually very happy with.
(3) The view is apparently widely held, and, interestingly, often on phenomenological grounds.
(4) “Out in the Silence” is a documentary that takes place, interestingly enough, about a half an hour down the road from where I grew up…

So far, empirical analyses of this phenomenon are largely lacking. A notable exception is Rohdenburg and Schlüter’s (2009) pilot study, which reveals a more advanced consolidation of evaluative sentence adverbs in British English, reflected in an American English tendency to rely on postmodification with *enough* to a greater degree. While semantically empty, *enough* reinforces the use of such adverbs as evaluative sentence adverbs and thus facilitates their processing, particularly since they instantiate reduced clauses whose actual subject, i.e. the speaker, is covert (cf. Swan 1991: 420; Lewis 2020: 25). In this regard, the addition of *enough* is

in line with Rohdenburg's (1996a) Complexity Principle and Mondorf and Pérez-Guerra's (2016) support strategies, i.e. the use of a functionally equivalent variant to reduce processing effort. Such support strategies should especially come into play with novel or weakly entrenched sentence adverbs (cf. Rohdenburg 1996b: 108).

Against this backdrop, the present paper pursues three aims. First, drawing on data from GloWbE, the study takes inventory of the attested range of *enough*-supported sentence adverbials, thus investigating to what extent British and American English differ in terms of type and token frequencies. This overview allows for revisiting some claims made by Schreiber (1971) concerning the alleged ungrammaticality of certain sentence adverbs, such as **impossibly* due to an Affect constraint, or the putative incompatibility of *enough* with modal sentence adverbs (**easily enough*). For both claims, the corpus provides counterevidence. A second aim is to zoom in on low-frequency types of *enough*-supported sentence adverbs, in particular hapax legomena constructions in both varieties, and to check whether respective counterparts without *enough* are attested, thereby providing further support for Rohdenburg's (1996b) assumption that *enough* conspicuously manifests with novel or weakly entrenched sentence adverbs. Third, in a follow-up study to Rohdenburg and Schlüter (2009), six high-frequency *enough*-supported sentence adverbs (*amazingly*, *astonishingly*, *curiously*, *interestingly*, *oddly*, *strangely*) are contrasted with their *enough*-less counterparts in order to analyse whether the data corroborate the British-American differences observed in their pilot study. A particular focus lies on the position of the sentence adverb within the clause, thereby differentiating initial, medial and final position, with the expectation that initial position, which has come to be the most established with sentence adverbs (cf. Swan and Breivig 2011), will attract *enough*-support the least.

References

Lewis, Diana (2020) Speaker stance and evaluative *-ly* adverbs in the Modern English period. In Kaltenböck, Gunther, María José López Couso & Belén Méndez Naya (eds.) *Investigating Stance in English: Synchrony and Diachrony*. *Language Sciences,* vol. 82 (Special issue). Online: https://www.sciencedirect.com/science/article/pii/ S0388000120300656

Mondorf, Britta & Javier Pérez-Guerra (2016) Support strategies in language variation and change. *Special Issue on Support Strategies in Language Variation and Change. English Language and Linguistics* 20(3): 383-393.

Rohdenburg, Günter (1996) Zur Einführung und Behauptung von lexikalischen Einheiten durch syntaktische Struktursignale im Englischen. In Weigand, Edda & Franz Hundsnurscher (eds.) *Lexical Structures and Language Use: Proceedings of the International Conference on Lexicology and Lexical Semantics, Münster, September 13-15, 1994*, 105-117. Tübingen: Niemeyer.

Rohdenburg, Günter & Schlüter, Julia (2009) New departures. In Rohdenburg, Günter and Julia Schlüter (eds.) *One Language, Two Grammars? Differences between British and American English*, 364-423. Cambridge: Cambridge University Press.

Schreiber, Peter (1971) Some constraints on the formation of English sentence adverbs. *Linguistic Inquiry* 2(1): 83-101.

Swan, Toril & Leiv Egil Breivik, (2011) English sentence adverbials in a discourse and cognitive perspective. *English Studies* 92(6): 679-692.

Swan, Toril (1991) Adverbial shifts: Evidence from Norwegian and English. In: Kastovsky, Dieter (ed.) *Historical English Syntax*, 409-438. Berlin/New York: Mouton de Gruyter.

He smiled unflinchingly: A corpus-based study on **ingly* adverbs as a reliable index of JK Rowling's authorship

Matthias Eitelmann and Ulrike Stange-Hundsdörfer (University of Mainz)

This study is concerned with the question of whether adverbs based on present participles like *smilingly, unerringly*, etc. truly are a characteristic of the *Harry Potter* series (as has been previously claimed by Broccias (2012), who therefore calls them "Harry Potter adverbs") or if they pattern similarly in other works of fiction. With Rowling publishing under the pen name Robert Galbraith, this question also concerns the issue of authorship attribution. Indeed, an earlier study on authorship that investigated whether Robert Galbraith is in fact JK Rowling, succeeded in identifying her as the most likely author based on variables such as the distribution of word length, the 100 most common words, bigrams and character 4-grams (Juola 2013).

The present study seeks to replicate Juola's findings focusing on **ingly* adverbs. The data used in the pilot study consists of four subsets: the *Harry Potter* series authored by JK Rowling (1.1m words), the *Cormoran Strike* series (written by JKR under the penname of Robert Galbraith; 645k words for volumes 1-3), the novel *The Casual Vacancy* (JKR; 160k words) as well as various pieces of crime fiction (written by PD James, Val McDermid, Ruth Rendell; 327k words). The analysis is based on all **ingly* adverbs as attested in the subsets (N=671), coded for function (with adjective and adverb modification expressing DEGREE, verb modification expressing CIRCUMSTANCE, and clause modification expressing STANCE, cf. Quirk et al. (1985) and Biber et al. (1999)) and verb semantics (MOTION, SPEECH, VISION, OTHER). Interestingly, **ingly* adverbs are *less* frequent in JKR's works than they are in the *Crime Corpus* (CC). They do differ significantly, however, with respect to their function ($p<0.001$, ctree analysis): verb modification (1) predominates in JKR's works (no matter the genre), while adjective (2), adverb (3) and clausal modification (4) are next to non-existent.

(1) She smiled encouragingly. (HP7)
(2) [H]e followed the achingly precise instructions […]. (CC)
(3) […] a hex that caused toenails to grow alarmingly fast. (HP6)
(4) Amazingly, he hadn't left sufficient forensic traces […]. (CC)

In the CC, on the other hand, both adjective and verb modification are common. Zooming in on verb modification reveals that the modification of SPEECH is very important across JKR's works, while there is an even distribution across the semantic categories in the CC ($p<0.001$, ctree analysis). Thus, the paper discusses to what extent *-ingly* adverbs serve as a reliable index of authorship particularly with respect to type-token ratios, functional distributions and the more or less creative implementation of the underlying word-formation pattern as attested in hapax formations. For the presentation, the *Cormoran Strike* dataset will also include Galbraith's volumes 4-7 as well as additional works in the *Crime Corpus* (target total: 1m tokens).

References

Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan & Edward Finegan (1999*) Longman Grammar of Spoken and Written English.* Essex: Longman.

Broccias, Cristiano (2012) Oriented *-ingly* adjuncts in Late Modern English. In Sauer, Hans & Gaby Waxenberger (eds.) *English Historical Linguistics 2008. Selected Papers from the Fifteenth International Conference on English Historical Linguistics*, 147-164. Amsterdam: Benjamins.

Juola, Patrick (2013) How a computer program helped show J.K. Rowling write A Cuckoo's Calling. Online: https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/ [14.10.2019]

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

Detecting cross-cultural differences in register variation across varieties of English

Stephanie Evert[1], Florian Frenken[2], Stella Neumann[2] and Gerold Schneider[3]

([1]FAU Erlangen-Nürnberg, [2]RWTH University of Aachen, [3]Univeristy of Zurich)

Registers reflect the constraints of systematically recurring situational contexts and are therefore embedded in the lingua-culture in which these situations arise. Consequently, when a language – such as English – is used in widely differing cultural contexts, the question arises whether registers in different varieties of the language might not actually reflect cultural differences between similar types of situations. Szmrecsanyi and Kortmann (2009) have shown that varieties of English fall into different clusters based on whether specific vernacular features are attested in the Electronic World Atlas of Varieties of English (Kortmann, Lunkenheimer, and Ehret 2020). Neumann's (2020) multivariate exploration of three components of the International Corpus of English (ICE; Greenbaum 1996) using a Geometric Multivariate Analysis (GMA) methodology indicates that (unsurprisingly) informal spoken texts in particular reflect differences between the varieties. In a revised replication of her GMA with a focus on register variation, Neumann and Evert (2021) suggest that register-related patterns of variation are much more pronounced than differences between varieties. However, they also observe divergence between texts in the same register from different varieties. The generality of both findings is limited, though, because their analysis was based on only three varieties of English.

Our paper aims at exploring these questions more thoroughly by drawing on a larger set of nine ICE components preprocessed for comparability (Lehmann and Schneider 2012) and by focusing the interpretation on registers that are expected to be more strongly affected by cultural differences. To this end, we extract the same set of 41 lexico-grammatical features from the ICE components as Neumann and Evert (2021), building on the CQP corpus queries (Evert et al. 2020) made available in their online supplement. Most of these queries rely on the rich part-of-speech tagset of the CLAWS tagger (Garside and Smith 1997).

Replicating the GMA analysis of Neumann and Evert (2021) allows us to first address some methodological gaps: based on the analysis of the full corpus we ask to what extent the results of Neumann and Evert (2021) depended on their specific choice of three ICE components. We will then focus on text categories that emerge clearly as conceptually spoken registers in the multivariate analysis (i.e. are close to Neumann and Evert's spoken pole of the "conceptual writing – conceptual speaking" dimension) in order to investigate cultural differences in situational contexts. Methodologically, we ask how latent dimensions change if only a subset of the text categories is included (especially focusing on conceptually spoken registers) and how different separate GMA analyses are for the three individual ICE components. Substantively, we explore whether the texts from one and the same spoken text category form clusters for the different varieties and ask whether multiple varieties combine into bigger clusters reflecting different variety types as would be predicted by Szmrecsanyi and Kortmann (2009).

The results are expected to shed light on the strength of the effect of cultural differences on registers given the well-documented robustness of register variation in multivariate studies of linguistic variation.

References

Evert, Stefan and The CWB Development Team (2020). The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial. CWB Version 3.5. http://cwb.sourceforge.net/documenta2on.php

Garside, Roger and Nicholas Smith. 1997. A Hybrid Grammatical Tagger: CLAWS4. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, edited by Roger Garside, Geoffrey Leech, and

Anthony McEnery, 102–21. London: Longman. http://ucrel.lancs.ac.uk/papers/HybridTaggerGS97.pdf.
Greenbaum, Sidney, ed. 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.
Kortmann, Bernd, Kerstin Lunkenheimer and Katharina Ehret, eds. 2020. *eWAVE - The Electronic World Atlas of Varieties of English*. Zenodo. 10.5281/zenodo.3712132.
Lehmann, Hans Martin and Gerold Schneider. 2012. BNC Dependency Bank 1.0. In *Studies in Variation, Contacts and Change in English 12. Aspects of Corpus Linguistics: Compilation, Annotation, Analysis,* edited by Signe Oksemell Ebeling, Jarle Ebeling and Hilde Hasselgård. Helsinki: VARIENG.
Neumann, Stella. 2020. On the Interaction between Register Variation and Regional Varieties in English. *Language, Context and Text* 2 (1): 121–44. https://doi.org/10.1075/langct.00023.neu.
Neumann, Stella and Stefan Evert. 2021. A Register Variation Perspective on Varieties of English. In *Corpus- Based Approaches to Register Variation*, edited by Elena Seoane and Douglas Biber, 143–78. Amsterdam/Philadelphia: John Benjamins.
Szmrecsanyi, Benedikt and Bernd Kortmann. 2009. The Morphosyntax of Varieties of English Worldwide: A Quantitative Perspective. *Lingua* 119 (11): 1643–63. https://doi.org/10.1016/j.lingua.2007.09.016.

Contrastive analysis of prepositional usage in German, English, Polish and Ukrainian: A corpus and dictionary-based approach

Iryna Fokashchuk and Peter Uhrig (FAU Erlangen-Nürnberg)

This study investigates the use of prepositions for expressing abstract relationships in German, English, Polish, and Ukrainian, positioning German as the source and the others as the target languages for comparative analysis. The aim is to discover if there is any cross-linguistic parallelism in NP1+P+NP2 pattern (e.g., *Interesse an jemandem/etwas, influence on somebody/something*), and to explore the potential reasons behind any similarities, utilizing the frameworks of Construction Grammar (CxG) and Conceptual Metaphor Theory (CMT).

The analysis is based on 120 examples of the NP1+P(auf/an)+NP2 structure, sourced from LGDaF (2019, 3rd ed.) along with the equivalent constructs in English, Polish, and Ukrainian. The corresponding constructs were searched for in the monolingual learner's dictionaries,* selected for their comprehensive lexicon entries, which encompassed governed prepositions, a detail often absent in bilingual dictionaries. For instance, in the Langenscheidt German-English online dictionary, the noun *Klage* is listed with *um* and *über,* but not *auf.* Contrastingly, Langenscheidt monolingual dictionary includes a more comprehensive entry, listing *über, auf* and *gegen* as governed by *Klage.* Similar gaps were observed in German-Polish and German-Ukrainian online dictionaries, such as PONS, Glosbe, and dict.cc. Furthermore, corpora such as DWDS (62 billion words) and DeReKo (55 billion) for German, COCA (1 billion) and BNC (100 million) for English, NKJP (1 billion) for Polish, and MOVA.info (100 million) for Ukrainian, were used to check the actual usage patterns in various contexts.

The result revealed an intriguing pattern: dictionaries in German, English, and to a significant extent in Polish, provided details on governed prepositions, which aligned with findings in the respective corpora. In contrast, this trend was not mirrored in Ukrainian. Frequently, the dictionary entries in Ukrainian lacked the governed prepositions, even though they were consistently observed in the corpus. This indicates that German, English, and Polish lexicography for nouns might be more comprehensive and potentially corpus-driven, unlike in Ukrainian.

Applying CMT, which suggests that we conceptualize abstract ideas based on more concrete ones (Lakoff and Johnson, 1980), to the gathered data yields an insightful observation: numerous examples demonstrate that abstract relationships are frequently expressed using spatial prepositions, albeit in distinct ways and not uniformly. For example, when German uses

auf, English often uses *to*, as in *Anspruch auf etwas* ('entitlement to something'). However, English, Polish, and Ukrainian often employ rather abstract prepositions such as *of, for, about* for relationships that German typically expresses using *auf* or *an*.

Polish and Ukrainian differ in their approach, with 26 of 120 instances not using any prepositions, especially in the scenarios where German employs *an*. For example, *Nachschub an etwas* in German corresponds to *dostaw czegoś* in Polish and *постачання чогось* in Ukrainian, whereas English consistently uses *of*.

In our presentation, we will give an extensive overview of the parallels and dissimilarities in expressing abstract relations across the four languages and the implications for linguistic theory.

*Note: *The Oxford Advanced Learner's Dictionary* (2020, 10th ed.) was used for the analysis of English, 'Inny słownik języka polskiego' (2000) was referred to for Polish, and 'Універсальний словник української мови' (2007) was consulted for Ukrainian.

References

Lakoff, George & Mark Johnson. 1980. *Metaphors we live by*. Chicago and London: University of Chicago Press.

Dictionaries

Bańko, Mirosław. 2000. *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa. Dict.cc. https://www.dict.cc/ [accessed November 2023]

Digitales Wörterbuch der deutschen Sprache (DWDS). https://www.dwds.de/wb/W%C3%B6rterbuch [accessed November 2023]

Glosbe. https://pl.glosbe.com/ [accessed November 2023]

Langenscheidt. 2019. *Langenscheidt Großwörterbuch Deutsch als Fremdsprache* (3rd ed.) Langenscheidt. [LGDAF]

Oxford Advanced Learner's Dictionary. https://www.oxfordlearnersdictionaries.com [accessed November 2023]

PONS. https://pl.pons.com/t%C5%82umaczenie/niemiecki-polski [accessed November 2023]

Куньч, Зорян. 2007. Універсальний словник української мови. Навчальна книга - Богдан.

Corpora

British National Corpus (BNC). https://www.english-corpora.org/bnc/

Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/ Das Deutsche Referenzkorpus (DeReKo). https://cosmas2.ids-mannheim.de/cosmas2-web/

Digitales Wörterbuch der deutschen Sprache (DWDS). https://www.dwds.de/d/korpora

MOVA.info. http://www.mova.info/corpus.aspx?l1=209

Narodowy Korpus Języka Polskiego (NKJP). http://nkjp.pl/

A multifactorial analysis of adjectives in Sri Lankan English

Nina Funke and Karola Schmidt (University of Giessen)

The present study investigates the adjective comparison alternation, i.e. the choice between synthetic and analytic adjective comparison, in Sri Lankan English (SLE). In recent years, the lexis-grammar interface of SLE has gained increased focus in corpus-linguistic research (cf. e.g. Bernaisch, 2015 on lexicogrammar; Götz, 2017 on fronting; Gries et al., 2021 on the genitive alternation; Hundt et al., 2020 on aspect). However, SLE has not yet been analysed in terms of adjective comparison as previous research has, so far, been limited to British English (BrE) language data (cf. e.g. Cheung & Zhang, 2016; Mondorf, 2003). These (often descriptive or monofactorial) studies identify the influence of, among others, length of the adjective, its

syntactic function, its frequency in the respective BrE corpora in question, and the presence of a complement as viable variables influencing this choice. Therefore, we understand our study as an addition to the growing body of research into the lexico-grammatical structures of SLE, as well as a further multifactorial exploration of the variety. To do this, we look into the alternation of the analytic or periphrastic adjective comparison, as exemplified in (1), and the synthetic or inflectional adjective comparison, as shown in (2), in present-day English.

(1) […] events that are most likely to happen, […] (SAVE2020 LK; 4719744)
(2) […] can be a bit harsher on the skin. (SAVE2020 LK; 4725069)

We extracted 446 data points, consisting of both comparative and superlative adjective forms, from the 2020 update of the South Asian Varieties of English (SAVE) corpus (cf. Bernaisch et al., 2021) and a selection of BrE data from 2020, as the historical predecessor of SLE, of the News on the Web (NOW2020) corpus (cf. Davies, 2016-) and seek to answer the following research questions:

- Does the choice of adjective comparison in SLE differ from BrE?
- What factors other than VARIETY influence the choice of comparison in BrE and SLE?

In a random forest analysis, we included the following predictors: ADJECTIVE_LENGTH, ADVERBIAL_MODIFICATION, COMPLEMENT, FORM, FREQUENCY, HAPLOLOGY, PERSISTENCE, RHYTHM, SEGMENT, STRESS_ON_LAST_SYLLABLE, SYNTACTIC_FUNCTION, and VARIETY well as all two-way interactions with VARIETY following a suggestion by Gries (2020). The random forest revealed that, unsurprisingly, the length of the adjective is the most important predictor in the model, in line with findings by Szmrecsanyi (2006) and Hilpert (2008). Both the presence or absence of stress on the last syllable of the adjective lemma and the variety of the speaker in interaction with other factors play an important role in our model, but neither factor has received as much attention as ADJECTIVE_LENGTH in previous multifactorial studies (with some notable exceptions, e.g. Szmrecsanyi, 2005 for stress).

References

Bernaisch, T. (2015). *The lexis and lexicogrammar of Sri Lankan English*. John Benjamins.

Bernaisch, T., Heller, B. & Mukherjee, J. (2021). *Manual for the 2020-update of the South Asian varieties of English (SAVE2020) corpus. Version 1.1*. Justus Liebig University, Department of English.

Cheung, L., & Zhang, L. (2016). Determinants of the synthetic–analytic variation across English comparatives and superlatives. English *Language and Linguistics* 20, 559–583. https://doi.org/10.1017/S1360674316000368

Sugar, spice and all things nice: A corpus-driven analysis of nominal and adjectival post-modification of English superordinate and light nouns

Sara Gesuato (University of Padova)

Most English noun phrases are characterised by a left-branching structure (*the big house; a do-it-yourself kit; world-shaking news; the special victims unit*). However, right-branching structures are also attested, e.g.: when the head noun includes a quantifier morpheme (*somebody experienced*, *something adventurous*), or is part of a loanword (*president elect*, *attorney general*), and also when the head noun is followed by an appositional expansion (*enemies, foreign and domestic*; *my life, long or short*), an adjective with complementation (*children*

interested in gardening; ideas worth considering), a name/title or identifying label (*owner Jane Smith*, *the film "The gladiator"*; *exercise B on page 5*), a generic adverb (*somebody else*; *no help whatsoever*) or a participle (*in solidarity with all people striking, all people concerned*).

This exploratory study examined the construction (Goldberg 1995, 2006) consisting of a noun post-modified by a noun or adjective with no complementation (e.g. *all things morphology*; *all matters American*) on the basis of corpus data. Using the CQL search function in the Sketch Engine platform, *English Trends*, a monitor corpus of mostly news articles, was searched for instances of 50 nouns denoting superordinate categories (Goddard 2017; e.g. *animals, furniture*) and plural light nouns (Simone, Masini 2014; Masini 2016; e.g. *things, ways*) preceded by "all" and followed by non-appositional nouns/adjectives, i.e. with no intervening punctuation marks. The chosen head nouns denoted abstract and concrete entities (e.g. *feelings*; *vehicles*), and animate, including human, beings (e.g. *creatures*; *individuals*); they comprised countable and mass nouns (e.g. *trends, time*), of Germanic and Latinate origin (e.g. *tools*, *utensils*), including derivatives (e.g. *publications*). The twofold goal was to determine which types of head nouns and which types of post-modifiers are co-selected for this construction.

The preliminary findings showed the following:

1) nominal post-modifiers are attested only with *matters* (74 tokens, 66 types) and *things* (78 tokens, 73 types), and include proper names (15 for *matters* and 8 for *things*);
2) only 4 of the nouns considered (*creatures, matters, things, ways*) have more than 50 instances of adjectival post-modification;
3) *matters* and *things* have the richest adjectival post-modification (43 types, 82 tokens for *matters*; 100 types, 10,000 tokens for *things*);
4) *matters* has mostly taxonomic adjectival post-modifiers (*verbal, financial, biographical*), while *things* has both taxonomic (*British*, *mobile, local*) and descriptive ones (*creative, cute, festive*), including those with a predicative function (*alive*, *onboard*), those serving as condensed relative clauses (i.e. [that are] *appropriate, available, imaginable, necessary*), and "binomial expressions" (e.g. *big and small*);
5) the most frequent adjectival post-modifiers are *available*, *possible,* and *great and small*.

The form-meaning pairing investigated is not fully predictable from its components: semantically, it is paraphrasable in various ways (e.g. 'things that are X, matters relevant to X'); formally, it is compatible with nominal and (descriptive and taxonomic) adjectival expansion, and frequently attested only with a few head nouns. The study suggests that this is a central construction, with variations on it extending from the "prototype".

References

Goldberg A. (1995) *Constructions: A Construction Grammar account of argument structure*. Chicago: University of Chicago Press.

Goldberg A. (2006) *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Masini F. (2016) Binominal constructions in Italian of the N1-di-N2 type: Towards a typology of Light Noun Constructions. *Language Sciences* 53: 99-113.

Sketch Engine (https://www.sketchengine.eu/)

Simone R. and Masini F. (2014) On light nouns. In Simone R. and Masini F. (eds.) *Word classes: Nature, typology and representations*, Amsterdam: John Benjamins, 51-74.

Vocabulary sophistication in primary schoolchildren's writing: A diachronic exploration

Victorina González-Díaz and Philip Durrant (University of Liverpool, University of Exeter)

Previous quantitative research on children's writing has explored learners' vocabulary development via interrelated constructs of lexical richness; the central idea being that a learner's lexicon becomes more *diverse* (includes a greater variety of types) and *sophisticated* as their writing matures throughout the school years. While lexical richness has been operationalised in different ways, it is the concept of sophistication that has standardly included a wider variety of measures, including register-based considerations such as the presence of academic vocabulary, the use of Greco-Roman lexis or the avoidance of informal (colloquial and non-standard) words (see references in Durrant, Brenchley and McCallum 2021; also Elliott et al. 2016; Constantinou and Chambers 2020.

As recent work by Durrant and associates indicates (Durrant and Brenchley 2019; Durrant and Durrant 2022), previous measures of sophistication only provide a limited picture of developmental complexities, as they align the notion of lexical sophistication with the specific vocabulary demands of the written academic register. They therefore propose a new register-based measure of sophistication, *appropriateness*, which focuses on a learner's ability to shape their vocabulary choices to the register they intend to invoke (Durrant and Durrant 2022: 51). Their studies also suggest that register-appropriate vocabulary development is discipline-specific and that – crucially for the present paper – lexical sophistication in school-writing is determined by the combination of two (so far considered separate) vocabulary measures: diversity and academic vocabulary use.

Taking Durrant and Durrant's (2022) methods and synchronic findings as starting point, this paper explores variation and change in vocabulary sophistication in the UK primary school context. The baseline of our analysis is the *Writing over Time* corpus, a recently-developed diachronic corpus of narrative and argumentative school writing (1979-2021; Merseyside area). Our paper addresses the following questions:

1) Are there any differences in vocabulary diversity in primary schoolchildren's writing across genres and time?
2) Does the use of register-appropriate vocabulary vary across time and genre in primary schoolchildren's writing?
3) What do the diachronic similarities and/or differences observed in (1) and (2) above tell us about vocabulary sophistication in school writing across time?

Our results consistently show a higher presence of more diverse and register-appropriate (academic) vocabulary in the modern (2021) Merseyside writing samples, although with some noticeable differences across genres: higher lexical diversity in the 2021 narratives is driven by greater use of general word-types, whereas in arguments the main difference lies in the greater presence of strongly-academic lexis in the 2021 data. Overall, the effect of these vocabulary differences is a more mature and consistent handling of register in the modern data. At a wider level, our findings do not align with previous diachronic analyses of schoolchildren's vocabulary sophistication (Constantinou et al 2019: 75ff; Constantinou and Chambers 2020) which record a gradual decrease in vocabulary diversity, and a parallel increase of register-*in*appropriate (colloquial and non-standard) lexis in secondary high-stakes examinations between 2004 and 2014. These discrepancies highlight the need for further attention to diachronic explorations of school writing development and their socio-educational implications, which this paper seeks to contribute to.

References
Constantinou, F. & Chambers, L. (2020). Non-standard English in UK students' writing over time. *Language and Education* 34(1), 22-35.
Constantinou, F., Chambers, L., Zanini, N. and Klir, N. (2019). A diachronic perspective on formality in students' writing: Empirical findings from the UK. *Language, Culture and Curriculum* 33(1), 66-83.
Durrant, P. & Brenchley, M. (2019). Development of vocabulary sophistication across genres in English children's writing. *Reading and Writing* 32(8), 1927–1953.
Durrant, P., Brenchley, M. & McCallum, L. (2021). Understanding development and proficiency in writing: Quantitative corpus linguistic approaches. Cambridge: Cambridge University Press.
Durrant, P. & Durrant, A. (2022) Appropriateness as an aspect of lexical richness: What do quantitative measures tell us about children's writing? *Assessing Writing* 51(2).
Elliott, G., Green, S., Constantinou, F., Vitello, S., Chambers, L., Rushton, N., Ireland, J., Bowyer, J. and Beauchamp, D. (2016). Variations in aspects of writing in 16+ English examinations between 1980 and 2014. *Research Matters: A Cambridge Assessment publication*. http://www.cambridgeassessment.org.uk/Images/340982-research-matters-special-issue-4- aspects-of-writing-1980-2014.pdf.

Exploring metadiscourse variations in learners' speaking and writing

Wenwen Guan and Bertus van Rooy (University of Amsterdam)

Metadiscourse (MD) denotes a rhetoric strategy that is used to highlight textual connections and actively engage the addressees in communication, no matter written or spoken. Having been thoroughly studied in written registers especially academic writing (e.g., Hyland, 2004; Li, 2012; Kim & Lim, 2013), research on MD has shifted towards spoken registers and contrasts between the two. A recent representative is Zhang's (2022) multidimensional study of variation among native speakers' English. It was discovered that MD, especially the interactive types, appears more in spoken registers than in written registers. Besides, she also reported that some scripted registers have more common MD usage with written registers than other conversational spoken registers. Comparative studies like this provide solid proof for MD as a highly register-sensitive rhetoric strategy. The diversified use of MD indicates language users' awareness of different communicative goals those registers are supposed to achieve. Additionally, based on the noticeable disparities in MD usage between native speakers and language learners that have been found in previous work (e.g., Lee & Deakin, 2016; Ädel, 2006), we are eager to investigate if learners also attempt to make a distinction among registers by observing the distribution of MD categories in their speaking and writing.

In this study, we selected Chinese learners' and native speakers' data of three registers, including spoken monologues, spoken dialogues, and written essays, from the International Corpus Network of Asian Learners of English (ICNALE). The raw data, namely 500,672 tokens, have been annotated with a MD scheme which combines Hyland's (2005) model and Ädel's (2010) model, the two commonly used taxonomies. MD usage is measured firstly in the term of categorical frequency. Our primary hypothesis follows Ädel's (2006) finding that learners tend to adopt a more transparent communicative strategy, which results in more MD. In addition, we also aim to examine if learners' use of MD displays different features across the three registers. These comparisons are accounted for by a mixed-effects model.

Beyond that, this study suggests that within-category diversity, namely the MD forms per category, should also be a stimulating indicator of MD usage as inspired by the vocabulary studies in corpus linguistics. It has been overlooked in existing research. However, we firmly believe it does not only reflect learners' overall language proficiency but also reveals their efforts to fulfill their communicative purposes. In order to evaluate the within-category variation, the

type-token ratio (TTR) of MD is computed. A hypothesis is that MD shows more diversity in writing as people are able to review the prior parts and avoid repetition. It is also assumed that native speakers use more varied metadiscourse markers than learners given their proficiency of the language. The findings of this analysis will be a pointer to easily acquired categories and attention-worthy ones from the perspective of language learning.

References

Ädel, A. (2010). *Just to give you kind of a map of where we are going*: A Taxonomy of metadiscourse in spoken and written academic English. *Nordic Journal of English Studies* 9(2), 69. https://doi.org/10.35360/njes.218

Ädel, A. (2006). *Metadiscourse in L1 and L2 English: Annelie Ädel*. John Benjamins.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Continuum.

Hyland, K. (2004). Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing* 13(2), 133–151. https://doi.org/10.1016/j.jslw.2004.02.001

Kim, L. C. & Lim, J. M.H. (2013). Metadiscourse in English and Chinese research article introductions. *Discourse Studies* 15(2), 129–146. https://doi.org/10.1177/1461445612471476

Lee, J. J. & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: International metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing* 33, 21–34.

Li, T. (2012). Metadiscourse repertoire of L1 Mandarin undergraduates writing in English: A cross-contextual, cross-disciplinary study. *Journal of English for Academic Purposes* 12.

Zhang, M. (2022). Variation in metadiscourse across speech and writing: A multidimensional study. *Applied Linguistics* 43(5), 912-933. https://doi.org/10.1093/applin/amac012

Genitive variation in spoken Late Modern English: A multivariate analysis of the Old Bailey Corpus

Stephanie Hackert[1], Sarah Potye[1] and Diana Wengler[2]

([1]LMU Munich, [2]University of Regensburg)

Despite being “the best researched of all syntactic alternations in English” (Rosenbach 2014: 215), genitive variation, i.e., the choice between the inflectional *s*-genitive (e.g., *the British Monarchy’s website*) and the periphrastic *of*-construction (e.g., *the website of the British Monarchy*), is surprisingly rarely studied in spoken corpora, let alone for historical periods of the language. We present a multivariate analysis of the genitive alternation in the Old Bailey Corpus (Huber et al. 2016), which spans the years between 1720 and 1913 and consists of trial proceedings containing over 24 million words. Having been taken down in shorthand, these proceedings are arguably as close to “real” spoken language as is possible for said period and in their sheer size constitute an invaluable resource for studying variation and change in Late Modern English.

We extracted more than 10,000 tokens of interchangeable genitives (cf. Hinrichs & Szmrecsanyi 2007: 446-7), annotating them for a number of structural factors relating to semantics, phonology, and processing and parsing. Specifically, we included possessor animacy, the semantic relation between possessor and possessum, the presence or absence of a word-final sibilant consonant in the possessor noun phrase, possessor definiteness, and the relative syntactic weight of possessor and possessum. We chose these factors because they have been investigated for the widest range of data sets and have consistently been found to powerfully influence the choice between the two genitive variants (cf. Rosenbach 2014: 252-62). The Old

Bailey Corpus being richly annotated for social variables, we also considered speaker gender, social class, role in the court setting (e.g., witness, judge, defendant, victim, lawyer), and decade, the latter enabling us to sketch developments in real time. In order to investigate the contribution of each factor to the variation observed, we employed mixed-effects models including the random factors of possessor and possessum head and speaker.

One of the most interesting results of our study is that the processing- and parsing-related factor of syntactic weight has a surprisingly weak effect, which, however, aligns well with findings from diachronic studies of more recent periods of the language, which have shown its impact to have grown continuously since the nineteenth century (cf. Wolk et al. 2013: 402; Hackert & Wengler 2022: 20). In sum, our study supplies an important missing link in the study of genitive variation in English and its historical development.

References

Hackert, Stephanie & Diana Wengler. 2022. "Recent grammatical change in postcolonial Englishes: A real-time study of genitive variation in Caribbean and Indian newswriting." *Journal of English Linguistics* 50, 3-38.

Hinrichs, Lars & Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 1, 437-474.

Magnus Huber, Magnus Nissel & Karin Puga. 2016. The Old Bailey Corpus 2.0, 1720-1913: Manual. https://fedora.clarin-d.uni-saarland.de/oldbailey/downloads/OBC_2.0_Manual%202016-07-13.pdf (November 28, 2023).

Rosenbach, Anette. 2014. English genitive variation – the state of the art. *English Language and Linguistics* 18, 215-262.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English. *Diachronica* 30, 382-419.

Collecting dialects: Institutional collections, community archives, and linguistic diversity

Ewan D. Hannaford and Marc Alexander (University of Glasgow)

Work-In-Progress

Diversifying the content of institutional collections is an increasingly salient concern within cultural studies and archival practice, aimed at redressing homogeneous and hegemonic social perspectives that have resulted from historic under-representation of diverse voices within mainstream cultural settings (Hyvärinen, 2020; Crilly & Everitt, 2021). These diversified collections are also of relevance to corpus linguists, since they ideally offer access to large quantities of thematically coherent data from a diverse range of sources, enabling analysis that strongly adheres to key corpus principles of representativeness, sampling, and balance (Baker & McEnery, 2015). Consequently, linguistically diverse archival collections potentially enable new corpus research, such as identifying key features of regional and social language varieties, examining their use/avoidance across different texts and contexts, and understanding unique discourses/genres produced by particular communities of language speakers.

However, the linguistic diversity of existing institutional collections in the UK is imperfect, with major UK archival collections currently being limited to predominantly documenting mainstream narratives and populations (Prescott & Hughes, 2018). To address this, the *Our Heritage, Our Stories* (*OHOS*) project is combining humanities and computer science expertise

to provide access to diverse community-generated digital content as part of the UK national collection, enabling search and comparison across materials held and generated by community archives from across the country, alongside existing institutional collections held at *The National Archives* in the UK. By enabling communities to tell their stories, in their words, the *OHOS* project is consequently opening up a wealth of new materials for linguistic research, by providing access to authentic samples of UK regional and social language varieties as they are used by and presented for their community of speakers, rather than as they may be standardised or translated for external audiences.

However, the incorporation of diverse linguistic varieties into existing, standardised collections also poses methodological challenges, with the linguistics and AI team on the *OHOS* project working to address how linguistic diversity in community-generated materials can be made compatible with existing frameworks and how linguistic diversity can best be situated and promoted within these settings. The proposed work-in-progress talk explores these challenges and the on-going approaches used on the project to overcome them, including challenges of dialect identification and integration, multilingual representation and cross-linking, and issues of accessibility and visibility. In working through these complexities, we simultaneously discuss the prospective benefits of reconfiguring institutional collections to include historically underrepresented language communities and new areas of corpus linguistic research being opened by these resources. As a result, through showcasing the linguistics work of the *OHOS* project, we demonstrate both the challenges and transformative potential of linguistically diverse collections for corpus studies, language preservation, and language equity.

References

Baker, P. & McEnery, T. (2015). Introduction. In P. Baker & T. McEnery (eds.), *Corpora and discourse studies* (pp. 1–19). Palgrave Macmillan. https://doi.org/10.1057/9781137431738

Crilly, J. & Everitt, R., (eds.) (2021). *Narrative expansions: Interpreting decolonisation in academic libraries*. Routledge. https://doi.org/10.29085/9781783304998

Hyvärinen, M. (2020). Toward a theory of counter-narratives: Narrative contestation, cultural canonicity, and tellability. In K. Lueg & M. Lundholt (eds.), *Routledge handbook of counter-narratives* (pp. 17-29). Routledge. https://doi.org/10.4324/9780429279713-3

Prescott, A. & Hughes, L. M. (2018). Why do we digitize? The case for slow digitization. *Archive Journal*, available at: https://www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization

---

POS annotation for Early Modern English at both ends of the scale: Wrangling the tags in Shakespeare's First Folio and EEBO-TCP

Andrew Hardie (Lancaster University)

The recently published dictionary of the language of the theatrical works of Shakespeare by Culpeper et al. (2023a, b) draws on multiple layers of corpus annotation-spelling regularisation, POS tagging, and lemmatisation – in the Enhanced Shakespearean Corpus (ESC). As well as Shakespeare's plays in the ESC:Folio corpus, in scale about a million tokens, comparative datasets include ESC:EEBO, a 296 million token subset of EEBO-TCP (Murphy 2019). The challenges of POS tagging Early Modern English (EModE) text using the CLAWS4 tagger (Leech et al. 1994) are common to both (see Rayson et al 2007), but otherwise the two corpora represent contrasting extremes in terms of the amount of careful manual work it was possible or desirable to apply to guaranteeing the quality of annotation in each.

This paper presents the stages by which a modified CLAWS for Shakespeare was developed through iterative manual work on (a) the spelling regularisation that underpinned POS analysis and (b) postediting of CLAWS output in the tractable ESC:Folio. The resources thus generated were applied with zero manual intervention to the much larger corpus.

The "patches" to the underlying CLAWS are of three kinds. First, changes to the C6 tagset schema address differences in the grammatical structure of EModE to contemporary English. Second, patches applied to the CLAWS lexicon fill in words which have become obsolete; are marginal now but were frequent then; exhibit different possible POS tags (or different probability profiles across those tags); or need special treatment due to expressing morphological categories – the primary, but not sole, case of this being second person singular verb agreement which in the modified tagset is a category as distinct as third singular agreement. So that these can be disambiguated by a Markov model trained on contemporary English, it is necessary to suppress the distinct tags until after the primary disambiguation stage, and then patch the novel analyses back into the CLAWS output. The non-probabilistic part of CLAWS, its "idiom tagging" recognising partially specified token-and-tag sequences, required similar patching to the lexicon, to block application of rules that are irrelevant to EModE or which assume idiomatisation/grammaticalisation of multiword expressions which cannot be assumed to have taken place prior to the 1590s/1610s. Similar domain-specific patches are required to allow the lemmatisation process to deal aptly with second singular forms.

As this framework was developed alongside the annotation, including full manual post-editing, of ESC:Folio, the resources are necessarily optimised to that dataset. By contrast, they could not be re-tailored for the drastically larger ESC:EEBO in any realistic timeframe. The effect on the larger corpus's annotation of being processed in a system built around a much different target data type – in terms of register and genre, among other external factors, but also in terms of the quality of faithful reproduction of the source documents – will be considered in light of (a) implications for the use of the data for the Encyclopedia and (b) implications for other purposes.

References

Culpeper, J., Hardie, A. & Demmen, J. (2023a) *The Arden Encyclopedia of Shakespeare's Language: Dictionary A-M*. Bloomsbury.

Culpeper, J., Hardie, A. & Demmen, J. (2023b) *The Arden Encyclopedia of Shakespeare's Language: Dictionary N-Z*. Bloomsbury.

Leech, G., Garside, R., Bryant, M. 1994. CLAWS 4: The tagging of the British National Corpus. *In Proceedings of the 15th International Conference on Computational Linguistics* (COLING 94), Kyoto, Japan, 622–628. http://ucrel.lancs.ac.uk/papers/coling1994paper.pdf10.3115/991886.991996

Murphy, S. 2019. Shakespeare and his contemporaries: Designing a genre classification scheme for Early English Books Online 1560–1640. *ICAME Journal* 43: 59–82. https://doi.org/10.2478/icame-2019-000310.2478/icame-2019-0003

Rayson, P., Archer, D., Baron, A., Culpeper, J., & Smith, N. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In M. Davies, P. Rayson, S. Hunston, & P. Danielsson (eds.), *Proceedings of Corpus Linguistics* 2007. http://ucrel.lancs.ac.uk/publications/CL2007/paper/192_Paper.pdf

Language, schmanguage: A corpus-based exploration of the semantics of English *shm*-reduplication

Stefan Hartmann and Tobias Ungerer

(Heinrich Heine University Düsseldorf, Concordia University)

Recent years have seen increased interest in the concept of "extravagance" (e.g. Petré 2017, Ungerer & Hartmann 2020, Baumann & Mühlenbernd 2022), and in extravagant morphology in particular (Eitelmann & Haumann [eds.] 2022). This paper presents a corpus-based study on what is arguably a prime example of an "extravagant" construction, namely *shm*-reduplication: a pattern in which a word is immediately repeated, but the initial consonant or consonant cluster is either replaced by /ʃm/, or /ʃm/ is added to the beginning of a word if it begins with a vowel (McCarthy & Prince 1996), as exemplified in (1) (from ENCOW16AX).

(1) And I did, and didn't actually say anything, just sort of nn uh yuh un uh. Language, schmanguage.

So far, research on *shm*-reduplication has mainly focused on its phonological properties (e.g. Nevins & Vaux 2003, Kołłątaj 2016, but see Mattiello 2013). The present study adds a semantic and a multimodal perspective. Drawing on the web corpus ENCOW16AX (Schäfer & Bildhauer 2012, Schäfer 2015), we use a semantic vector-space analysis (e.g. Levshina & Heylen 2014, Perek 2016) to take a closer look at the semantic domains to which the base words in the construction belong. We show that the major semantic domains from which the instances of the pattern are drawn include law, education, and official institutions as well as health and food. This indicates that *shm*-reduplicatives owe at least some of their extravagant effect to the fact that they cast an ironic or sarcastic perspective on "serious" domains of everyday life. In addition, we test two hypotheses based on multimodal data from the TV News Archive. Firstly, we show that the construction is often accompanied by a dismissive gesture such as a member of the *away* gesture family (Bressem & Müller 2014), thus qualifying as a multimodal construction in the sense of e.g. Zima (2014). Secondly, we predicted that the construction tends to occur turn-initially, usually taking up cues from the interlocutor's previous utterance if occurring in a conversation. Here, however, the data show that matters are more complex. In fact, *shm*-reduplication tends to occur in what could be called "fictive quotes" (Pascual 2014): An attitude ascribed to a person is conveyed by a quote attributed to said person, as in (2).

(2) donald trump says debate shmebate. (Action News at 6:00 AM)

In many ways, then, the pattern is a typical example of a creative and "extravagant" construction that is strongly connected to specific communicative contexts and characterized by a fairly complex set of discourse-functional properties. A closer investigation of this and other expressive constructions can inform our understanding of the grammatical as well as the social and interactional aspects that underlie creative and playful language use. In addition, the multimodal perspective taken here can help us understand how gestures emphasize and enhance the "extravagant" character of such constructions.

References

Baumann, Andreas & Roland Mühlenbernd. 2022. Less of the same: Modeling horror aequi and extravagance as mechanisms of negative frequency dependence in linguistic diversification. In Andrea Ravignani, Rie Asano, Daria Valente, Francesco Ferretti, Stefan Hartmann, Misato Hayashi, Yannick Jadoul, Mauricio Martins, Yohei Oseki, Evelina Daniela Rodrigues, Olga Vasileva & Slawomir Wacewicz (eds.), *The evolution of language. Proceedings of the joint conference on language evolution (JCoLE)*, 50–57. Nijmegen: Max Planck Institute for Psycholinguistics.

Bressem, Jana & Cornelia Müller. 2014. The family of Away gestures: Negation, refusal, and negative assessment. In Cornelia Müller, Alan J. Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill & Sedinha Tessendorf (eds.), *Body - language - communication: An international handbook on multimodality in human interaction*, 1592–1604. Berlin, Boston: De Gruyter Mouton.

Eitelmann, Matthias & Dagmar Haumann (eds.). 2022. *Extravagant morphology*. Amsterdam, Philadelphia: John Benjamins.

Kołłątaj, Andrzej. 2016. Reduplication in English - Typology, correlation with slang and metaphorisation. *Philolog*. *Studia Neofilologiczne* 6, 237–248.

Levshina, Natalia & Kris Heylen. 2014. A radically data-driven Construction Grammar: Experiments with Dutch causative constructions. In Ronny Boogaart, Timothy Colleman & Gijsbert Rutten (eds.), *Extending the scope of Construction Grammar.* Berlin, New York: De Gruyter.

Mattiello, Elisa. 2013. *Extra-grammatical morphology in English. Abbreviations, blends, reduplicatives, and related phenomena*. Berlin, Boston: De Gruyter.

McCarthy, John J. & Alan Prince. [1986] 1996. *Prosodic morphology*. Final revision 1996. Linguistics Department Faculty Publication Series 13. https://scholarworks.umass.edu/linguist_faculty_pubs/13 [07.09.2018]

Nevins, Andrew & Bert Vaux. 2003. Metalinguistic, shmetalinguistic: the phonology of shm- reduplication. *Proceedings of the annual meeting of the Chicago Linguistic Society* 39(1), 702–721.

Pascual, Esther. 2014. Fictiv*e interaction: the conversation frame in thought, language, and discourse*. (Human Cognitive Processing 47). Amsterdam, Philadelphia: John Benjamins.

Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony. A case study. *Linguistics* 54(1), 149–188.

Petré, Peter. 2017. The extravagant progressive: an experimental corpus study on the history of emphatic be V-ing. *English Language and Linguistics* 21(2), 227–250. https://doi.org/10.1017/S1360674317000107.

Schäfer, Roland. 2015. Processing and querying large corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen & Andreas Witt (eds.), *Challenges in the management of large corpora (CMLC-3)*, 28–34.

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Cicoletta Calzolari, Khalid Choukri, Terry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of LREC 2012*, 486–493.

Ungerer, Tobias & Stefan Hartmann. 2020. Delineating extravagance: Assessing speakers' perceptions of imaginative constructional patterns. *Belgian Journal of Linguistics* 34, 345–356. https://doi.org/10.1075/bjl.00058.ung.

Zima, Elisabeth. 2014. Gibt es multimodale Konstruktionen? Eine Studie zu [V(motion) in circles] und [all the way from X PREP Y]. *Gesprächsforschung* 15, 1–48.

Corpora

ENCOW16AX = Corpora from the web, https://www.webcorpora.org; see Schäfer & Bildhauer (2012)

TV News Archive = https://archive.org/details/tv

Exploring grammatical patterns of expertise in Reddit discussions

Turo Hiltunen (University of Helsinki)

Public debates are increasingly conducted on various internet forums and social media platforms, and such sites of social interaction are good sources for studying language use in the context of scientific and technical communication and decision-making. This paper investigates grammatical patterns related specifically to experts and expert knowledge in online discussions about science and technology topics on the discussion forum *Reddit*. What makes expertise a particularly relevant topic in this context is the claim that the authority of expert knowledge and trust in experts is increasingly challenged in public forums, leading to a "crisis of expertise"

(Davies 2018, Eyak 2019), which some writers have attributed the internet (e.g. Nichols 2017). Building on earlier linguistic work on expert discourses (e.g. Hiltunen 2023) and on epistemic argumentation on Reddit (Dayter and Messerli 2022, Biri 2022), the present analysis addresses the following questions: (1) what grammatical patterns are used when referring to expert knowledge in Reddit discussions on science and technology topics, (2) how their frequency of use varies across different subregisters, and (3) what pragmatic and rhetorical functions such patterns typically have.

The analysis is based on data from the *Reddit Expertise Corpus*, a purpose-built 60-million-word dataset compiled from a total of 77 subforums (known as "subreddits") which are grouped under three broad topics: Science, Politics, and Questions and answers. Given its size of and structure, the corpus enables the quantitative analysis of how the patterns are used across subregisters in the data.

The identification and classification of grammar patterns is based on the analysis of concordances of core lexemes denoting experts and expert knowledge (e.g. *expert*, *scientist*; *survey*, *statistics*), and their functions are described in terms of data-driven *local grammars* (Hunston and Su 2019).

The analysis identifies several recurrent grammatical patterns related to expertise, which are primarily used for backing one's own arguments and discrediting those of others, but also for other functions. It is argued that from a sociological perspective, these patterns are typically concerned with *meta-expertises*, or the judging and choosing between substantive experts (Collins and Evans 2017). Alongside these frequent patterns, the data also exhibits a number of context-specific patterns and their associated local grammars. This in turn suggests that the use of expertise as a rhetorical strategy is closely related to register considerations and the norms of individual discourse communities. Overall, the corpus-based analysis of how experts and expert knowledge are framed and represented in the data contributes to the description of roles, contexts and conceptualisations of expertise in public discourses, which has previously been identified as an under-researched area in expertise studies (Conway and Gore 2019).

References

Biri, Ylva. 2022. Epistemic stance in the climate change debate: A comparison of proponents and sceptics on Twitter and Reddit. *Register Studies* 4:2, 232-262.

Collins, Harry and Robert Evans. 2007. *Rethinking expertise*. Chicago: University of Chicago Press.

Davies, William. 2018. *Nervous states. How feeling took over the world*. London: Jonathan Cape.

Dayter, Daria and Thomas C. Messerli. 2022. Persuasive language and features of formality on the r/ChangeMyView subreddit. *Internet Pragmatics* 5:1, 165–195.

Eyal, Gil. 2019. *The crisis of expertise*. Polity Press.

Hiltunen, Turo. 2023. This job requires considerable expertise: Tracking *experts* and *expert knowledge* in the British parliamentary record 1800–2005. In: Minna Korhonen, Haidee Kotze, and Jukka Tyrkkö (eds.), *Exploring language and society with big data: Parliamentary discourse across time and space*. Amsterdam: John Benjamins. 227–249.

Hunston, Susan and Hang Su. 2019. Patterns, constructions, and local grammar: A case study of 'evaluation'. *Applied Linguistics* 40:4, 567–593.

Nichols, Tom. 2017. *The death of expertise: The campaign against established knowledge and why it matters*. Oxford: Oxford University Press.

To clash or not to clash with thirteen men: The linguistic context of stress shift in authentic speech

Sebastian Hoffman and Sabine Arndt-Lappe (University of Trier)

It is a well-established fact that languages have rhythmic properties. For English, there is a considerable body of research on what has been termed the Principle of Rhythmic Alternation ('PRA', Sweet 1876) – i.e. the general tendency to maintain an alternation of stressed and unstressed syllables. The bulk of this research is on written data (or on orthographically transcribed speech), but a small number of studies have also investigated the phenomenon on the basis of authentic speech (Shattuck-Hufnagel et al. 1995, Azzabou-Kacem 2018, Arndt-Lappe and Hoffmann 2022). This type of work has focused on phonetic contexts that are known to trigger the so-called '*thírteen mén* rule', i.e. adjectives (or numerals) whose citation form is stressed on the word-final syllable occur before a noun that is stressed on its initial syllable (as in the combination of *thirtéen* and *mén*). In these contexts, stress shift on the prenominal adjective is commonly observed (as in *thírteen mén*) and is traditionally interpreted as a repair strategy to maintain rhythmic alternation.

The findings reported in Arndt-Lappe and Hoffmann (2022) confirmed that prenominal adjectives (or numerals) in English indeed show stress patterns that are compatible with the PRA, but that there is also a great deal of variation. In addition, the probability of stress shift was shown to be correlated with the prenominal token frequency of the adjective, suggesting that we may in fact not be dealing with stress shift in the first place. Instead, it may be the case that at least the frequent adjectives are retrieved in their 'shifted' form during speech production. In other words, this questions the idea that stress shift happens in real time in online processing.

For our follow-up study, we will again make use of the spoken component of the first *British National Corpus* (released in 1994). For a sizeable proportion of the corpus, audio recordings and a phonemic transcription are available (see Coleman et al. 2012), making it possible to retrieve potential stress clash patterns and to analyse their actual realisations in connected speech.

Our investigation focuses on all adjective-noun combinations in which the noun has lexical stress on the initial syllable and in which the adjective allows for stress shift. All 2,126 relevant observations were subjected to an acoustic analysis of the F0 contour of the adjective-noun combination and their surrounding phrasal context, making it possible to determine whether the initial syllable of the noun is accented – as would be expected by their citation form – or whether it is in fact deaccented in context. Preliminary findings suggest that in a sizeable proportion of cases, the noun is in fact deaccented. In other words, there is no potential for stress clash that could be avoided in the first place. Interestingly, we still find "shifted" constellations in such contexts, offering clear support for our interpretation that the concept of "stress shift" as such must be questioned. We will present a multivariate analysis of our data and discuss the theoretical implications of our findings.

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.

Azzabou-Kacem, Soundess. 2018. *Stress Shift in English Rhythm Rule Environments: Effects of Prosodic Boundary Strength and Stress Clash Types*. Doctoral thesis. Edinburgh: University of Edinburgh.

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.

Shattuck-Hufnagel, Stefanie, Mari Ostendorf and K. Ross. 1994. Stress shift and early pitch accent placement in lexical items in American English. *Journal of Phonetics* 22 (4). 357–88.

Sweet, Henry. 1876. Words, logic, and grammar. *Transactions of the Philological Society*, 1875– 1876. 470–503.

Towards a pipeline approach to corpus compilation: Challenges and solutions

Samuel Hollands and Hanna Schmück (University of Sheffield, Lancaster University)

Despite enormous financial and time investments in many modes of corpus compilation there is often a lack in consistency with regards to corpus processing, format, and structure (Demmen, 2020; Diemer et al., 2016, Reppen, 2022). Issues spanning domains and impacting comparability of spoken and written corpora (Lindquist & Levin, 2000) include irregular metadata formats, inconsistent data structures, varied transcription approaches, amongst others. In the domain of speech corpora, we see additional issues such as methodologically unjustified variation in audio formats. This paper aims to explore ways in which both written and spoken corpus compilation can be streamlined and, as an example, which best practices can be employed for constructing eBook corpora. Recommendations relevant to researchers working with spoken and written corpora are provided in order to highlight the importance of working towards a methodological conversion in these two domains.

Within this study we are proposing the Python Corpus Pipeline (PCPi) to streamline corpus compilation via programmatic blueprints for ideal corpus structures. This allows researchers to automatically format corpora into a regular schema such as XML or TEI (TEI consortium, 2023) and encourages conscious decisions in the early stages of corpus compilation. The objective is to provide a practical tool that helps researchers implement best practices and adhere e.g. to the FAIR principles (Wilkinson et al., 2016) in their workflow. As part of PCPi, specific recommendations are made to address issues in the spoken corpus domain such as varied microphone setups, irreversible acoustic post-processing of recordings, and inconsistent use of audio filetypes. For instance, low-bitrate MP3 – widely accepted for efficient signal compression and used in spoken corpora – aims to preserve human perception while removing audio information (Watkinson, 2012: 169-227). However, this compression can pose challenges for speech analytics and emotion recognition due to information loss (Campbell, 2002; Lotz, 2017).

In the written domain we provide a worked example from the currently ongoing compilation of the Lancaster-Northern Arizona Corpus of American English (LANA). Even straightforward tasks such as removing the front and backmatter pose significant problems when working with *epub* files that do not follow a rigorous standard. This is the case since long acknowledgements or reading samples run the risk of mimicking the style of the desired main body of the text and chapter breaks are not reliably marked. In our case study, the Fiction section of LANA, only 664 out of 1325 books available (50.1%) contain explicit and reliable chapter breaks due to significant formatting inconsistencies. A PCPi subroutine splits individual files into paragraphs and checks their contents in several passes to classify them as belonging to the main body of the text or not using a window-based approach, resulting in 1133 (85.5%) salvageable eBooks.

Beyond specialised applications such as text extraction from eBooks, PCPi generally streamlines corpus processing by ingesting raw text and XML data, applying NLP-driven enrichment, and interfacing with SpaCy and Pymusas for tasks like POS tagging, tree parsing, and named entity recognition. The framework accommodates varied XML structures while ensuring reliable downstream processing.

References

Campbell, N. (2002). Recording and storing of speech data. In *Proceedings LREC* (Vol. 7, pp. 109-114).

Demmen, J. (2020). Issues and challenges in compiling a corpus of Early Modern English plays for comparison with those of William Shakespeare. *ICAME Journal* 44(1), 37-68. https://doi.org/10.2478/icame-2020-0002

Diemer, S., Brunner, M.-L. & Schmidt, S. (2016). Compiling computer-mediated spoken language corpora. In *Compilation, transcription, markup and annotation of spoken corpora* (Vol. 21, Issue 3, pp. 348–371). John Benjamins Publishing Company. https://doi.org/10.1075/ijcl.21.3.03die

Lindquist, H. & Levin, M. (2000). Apples and oranges: On comparing data from different corpora. In *Corpus Linguistics and Linguistic Theory* (pp. 201–213). BRILL. https://doi.org/10.1163/9789004490758_017

Lotz, A. F., Siegert, I., Maruschke, M. & Wendemuth, A. (2017). Audio compression and its impact on emotion recognition in affective computing. In *Konferenz Elektronische Sprachsignalverarbeitung* (pp. 1-8). Dresden.

Reppen, R. (2022) Building a corpus: what are key considerations? In A. O'Keeffe & M.J. McCarthy (eds.) *The Routledge handbook of corpus Linguistics* (pp. 48–61). Routledge. doi:10.4324/9780367076399-5

TEI Consortium (2023). TEI P5: Guidelines for electronic text encoding and interchange. 4.7.0.16.11.2023. http://www.tei-c.org/Guidelines/P5/ (07.03.2024).

Watkinson, J. (2012). *The MPEG handbook*. Routledge.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*(1). https://doi.org/10.1038/sdata.2016.18

A key feature analysis of linguistic themes in award-nominated screenplays

Alexander Holmberg and Michael Edens

(Northern Arizona University, Montana State University)

Work-In-Progress

Telecinematic texts (e.g., movie and TV scripts) have garnered increased research interest in recent years (McIntyre, 2012). Findings from these studies have shown that dialogue in telecinematic texts tends to be fundamentally different from spoken conversation (Bednarek et al., 2021), thus establishing telecinematic texts as their own register. Furthermore, research into movie screenplays has shown that certain linguistic features (e.g., verb-pronoun clusters, noun clusters) differ between the two subregisters of screenplays: dialog and stage directions (scene and setting description) (Buckland, 2023). However, while the register characteristics and format of screenplays tend to remain constant across movies, the way movies are perceived does not: whereas some movies receive critical acclaim, others are heavily criticized. Although there are numerous possible factors at play that may explain this, one possibility is that there are reasons that can be captured through linguistic methods.

The present study aims to investigate the extent to which linguistic features seem to be associated with the critical reception of movie screenplays. Critical reception is operationalized with the help of two screenplay awards: the Academy Awards (Oscars) and the Golden Raspberry Awards (Razzies). The Oscars is the most renowned award for Best Screenplay, while the Razzies awards the Worst Screenplay award. This study addresses the two following research questions:

1) What lexico-grammatical features (e.g., attributive adjectives, technical nouns, discourse particles) are more commonly found in Oscar-nominated contemporary drama screenplays compared to Razzie-nominated contemporary drama screenplays?
2) To what extent may functional analysis of groupings of these features help provide possible explanations for the screenplay reception?

To answer these research questions, we compiled a corpus of screenplays: the Popular Corpus of Oscar and Razzie Nominated Screenplays (PopCORNS) corpus. It covers a total of 39 screenplays, which corresponds to 10% of the total number of Oscar and Razzie-nominated screenplays from 1992-2022.

We used Key Feature Analysis (Egbert & Biber, 2023) to identify what linguistic features differed between the two groups of screenplays. Key Feature Analysis measures key features between corpora or sub-corpora using Cohen's *d*. The threshold for a key feature in this study was set at *d* = .90 to best represent the most influential key features in each group.

Our results yielded 30 key features. These features were then categorized into seven functional groups: features of description, nouns, pronouns, attitudinal/communicative features, verb aspect/tense, discourse particles/connectivity/clausal features, and likelihood/modality features. Three main themes emerged: (i) Oscar-nominated screenplays contained more key features related to people. Razzie-nominated screenplays had more key features pertaining to things or events; (ii) Oscar-nominees had more key features relating to detailed yet concise stage directions compared to Razzie-nominees; (iii) Oscar-nominees had key features often prescribed to authentic conversational English. In sum, these initial findings indicate that there are notable linguistic differences between Oscar and Razzie-nominated screenplays, thus suggesting that there are indeed linguistic clues that may help us better understand perceived screenplay quality.

References

Bednarek, M., Pinto, M. V. & Werner, V. (2021). Corpus approaches to telecinematic language. *International Journal of Corpus Linguistics* 26(1), 1–9.

Buckland, W. (2023). The motion picture screenplay as data: Quantifying the stylistic differences between dialogue and scene text. In Davies, R., Russo, P. & Tieber, C. (eds.) *The Palgrave Handbook of Screenwriting Studies*. Palgrave Macmillan,

Egbert, J. & Biber, D. (2023). Key feature analysis: A simple, yet powerful method for comparing text varieties. *Corpora* 18(1), 121-133.

McIntyre, D. (2012). Prototypical characteristics of blockbuster movie dialogue: A corpus stylistic analysis. *Texas Studies in Literature and Language* 54(3), 402-425.

The effect of accuracy on grading of Swedish EFL students' writing during high-stakes exams

Christian Holmberg Sjöling (Luleå University of Technology)

Teaching students to express themselves accurately in writing is a part of teachers' everyday lives. Accuracy can function as an indicator of different stages of learners' language development and, thus, help teachers determine which linguistic features are difficult for students. While there are many different definitions of accuracy, researchers tend to agree that it concerns the amount of control learners have over a language system, defined as "freedom from errors" (Foster & Skehan 1996, pp. 196–197; Thewissen & Anishchanka 2022, pp. 211). The importance of linguistic accuracy should not be disregarded from a research perspective given

that both summative and formative English as a Foreign Language assessment practices consider both quantity and severity of errors when grading texts (Pallotti, 2010, pp. 159). This is the case in Swedish upper secondary school as students are required to take the National Tests of English every year to ensure that their proficiency is on par with the level at which they study (Olsson, 2018). During the tests, students are required to write texts on a specific topic and these are then assessed by teachers with assessment instructions created by a group of experts on behalf of the Swedish National Agency for Education. These instructions specify accuracy as a grading criterion and indicate that there should be a progression in terms of accuracy between the lowest and highest grade. Furthermore, to ensure fair and equal assessment, the instructions provide graded example texts to assist teachers' assessment.

This paper aims to examine the effect of accuracy on grading during the National Tests of English by analysing a corpus consisting of 142 graded example texts (50,048 words) and 175 teacher graded texts (76,924 words) written and assessed between 2011 and 2022. The answers to the following two research questions are sought: Which category of errors has the strongest impact on grading? Is accuracy assessed differently by expert raters and teachers? To do so, the material was manually annotated for errors with the newly released *Université Catholique de Louvain Error Editor* (Granger et al., 2023) and its default tagset the *Louvain Error Tagging Manual, version 2* (Granger et al., 2022). Then, the error-annotated data was quantified using Potential Occasion Analysis (Thewissen, 2021) before statistical analysis using ordinal forests and conditional inferences trees was carried out in *R* (Gries, 2023; Hornung, 2020). The preliminary results suggest, in both sets of texts, that form errors (spelling and morphological errors) have the strongest impact on grading followed by lexical errors for single verbs (e.g., *make the Christmas tree* instead of *dress the Christmas tree*), and grammar errors (particularly subject-verb agreement errors). There is also a discrepancy between expert raters and teachers as the former appear to be more lenient in their assessment practice. The findings are discussed in relation to assessment, test construction and their pedagogical implications.

References

Foster, P. & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition* 18(3), 299–323.

Granger, S., Swallow, H. & Thewissen, J. (2022). *The Louvain Error Tagging Manual. Version 2.0.* CECL Papers 4. Louvain-la-Neuve: Centre for English Corpus Linguistics/Université Catholique de Louvain. (38 pages).

Granger, S., Thewissen, J. & Swallow, H. (2023). *The UCLouvain Error Editor User Guide - Version 2.0.* CECL Papers 6. Louvain-la-Neuve: Centre for English Corpus Linguistics/Université Catholique de Louvain. (33 pages).

Gries, S. T. (2023). *Statistics for Linguistics with R: A Practical Introduction*, 3rd ed. Walter de Gruyter.

Hornung, R. (2020). Ordinal forests. *Journal of Classification 37*, 4–17.

Olsson, E. (2018). Vokabulär och bedömning av skriftlig förmåga. In G. Erickson (ed.) *Att bedöma språklig kompetens* (pp. 139–167). University of Gothenburg.

Pallotti, G. (2010). Doing interlanguage analysis in school contexts. *EUROSLA Monographs Series* 1, 159–190.

Thewissen, J. (2015). *Accuracy across proficiency levels: A learner corpus approach* (Vol. 2). Presses universitaires de Louvain.

Thewissen, J. (2021). Accuracy. In N. Tracy-Ventura and M. Paquot (eds.) *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 305–317). Routledge

Thewissen, J. & Anishchanka, A. (2022). Interaction between grammatical accuracy and syntactic complexity at different proficiency levels. In A. Leńko-Szymańska & S. Götz (eds.) *Complexity, Accuracy and Fluency in Learner Corpus Research* (pp. 209-240). Benjamins.

The long history of shortening: A diachronic analysis of abbreviation practices from the fifteenth to the twenty-first century

Alpo Honkapohja and Imogen Marcus (Tallinn University, Edge Hilll University)

Abbreviating words instead of spelling them out in full is a phenomenon found throughout the history of written communication, from ancient inscriptions carved into stone to modern-day instant messages. The paper seeks to bridge the gap between studies of earlier abbreviation practices and those present in CMC (computer-mediated communication) by using a corpus-based, long diachronic approach, data from the fifteenth to the twenty-first century, and a framework that examines both abbreviation types and lexis that is abbreviated. In so doing, it aims to lay the foundation for further diachronic studies of abbreviation practices.

Adopting a diachronic perspective and lexicological framework, it quantitatively analyses interlocking corpora in registers related to speech-like registers across fifteenth-century memoranda, letters and administrative receipts, seventeenth-century letters and depositions, late nineteenth-century letters, early twentieth-century letters and a subcorpus of WhatsApp instant messages dating from 2018–19.

Time period	Sub-corpus	Text type	Words
ME (1066-1500)	Middle English Local Documents Corpus (MELD)	15th-cent. letters	3,323
ME	Middle English Local Documents Corpus (MELD)	15th-cent. statements, receipts	1,705
ME	Middle English Local Documents Corpus (MELD)	15th-cent. memoranda	11,458
EModE (1500-1700)	The Corpus of Early English Correspondence (CEEC)	17th-cent. letters	21,580
EModE	English witness depositions 1560-1760: an electronic text edition (ETED)	17th-cent. depositions	4,291
LModE (1700-1945)	The Corpus of Early English Correspondence (CEEC)	Late 19th-cent. letters	14,456
LModE	Corpus of Late Modern English Prose, Project Gutenberg, Imperial War Museums Website	Early 20th-cent. letters	11,427
21st cent. English (2000-present)	Transhistorical Corpus of Written English (TCWE)	21st-cent. instant messages	21,228

We collected a dataset of abbreviated spellings in each subcorpus and annotated them for both **abbreviation form** (e.g. brevigraph, contraction, clipping, superscript) and **lexeme category** (e.g. name, title, function word, expression of time). This dataset was then subjected to exploratory quantitative analyses, including descriptive statistics (specifically log likelihood tests). We also carried out a qualitative analysis of these lexeme categories over the centuries, with a focus on specific examples.

Major changes to overall abbreviation density across time are identified. The forms of abbreviation also go through major change, but the types of lexemes that are abbreviated stay more consistent over time. For example, abbreviations being used for closed-class function words such as *the* and *that* are dominant from the earliest data we have looked at to the present day. Overall, the study demonstrates how situating new media abbreviation practices within a historical continuum can enhance our understanding of them.

References
Bieswanger, Markus. 2013. Micro-linguistic structural features of computer-mediated communication. In Herring, Susan, Stein, Dieter & Virtanen, Tuija (eds.), *Pragmatics of computer-mediated communication*, 463–85. Berlin: De Gruyter Mouton.
Cannon, Garland. 1989. Abbreviations and acronyms in English word-formation. *American Speech* 64(2), 99– 127.
González Rodríguez, Félix & Cannon, Garland. 1994. Remarks on the origin and evolution of abbreviations and acronyms. In Moreno Fernández, Francisco, Fuster, Miguel & Calvo, Juan José (eds.), *English historical linguistics* 1992, 261–72.

Mapping the comparative correlative across the GloWbE: More evidence for constructional networks

Jakob Horsch (Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences)

The English Comparative Correlative (CC) is a bi-clausal construction ([*The more I read*]$_{C1}$ [*the more I know*]$_{C2}$) that has attracted interest due to its semantics and formal features. It has mostly been described in the context of introspective studies that tried to account for it with maximally abstract rules/templates (e.g. Culicover and Jackendoff 1999, Borsley 2004, den Dikken 2005). However, recent corpus studies (Hoffmann et al. 2019, 2020) have provided evidence for statistically significant syntactic inter-dependencies between C1 and C2 that Hoffmann et al. refer to as "cross-clausal associations" (2019: 32). These cannot be modeled with maximally abstract rules and templates, which is why Hoffmann et al. have suggested a Usage-based Construction Grammar (CxG) approach, assuming that an inheritance network of interrelated constructions underlies the CC construction (2019).

While insightful, Hoffmann et al.'s studies were limited to the standard varieties British English (2020) and American English (2019). Therefore, the question remains whether the many other varieties of English have similar networks. This was to be expected, following Goldberg's Tenet #5 that predicts cross-linguistic generalizations as a result of "general cognitive constraints" (2003: 219). This includes the entrenchment of constructions as a result of "sufficient frequency" (Goldberg 2006: 5), which in turn can be determined using corpus data. A further question was what the constructional networks in the non-standard varieties look like. Some degree of variation was expected, since domain-general cognitive processes such as chunking and repetition (Bybee 2012, Croft 2013: 224) can lead to the entrenchment of language-specific idiosyncrasies. Following Goldberg's Tenet #6 (2003: 219), these can also be captured by inheritance networks.

To address this 'blind spot' on the world map of English, I conducted a corpus study based on over 5,500 CC tokens from the GloWbE corpus covering 20 varieties of English. To facilitate data analysis and exposition of results, I conflated the 20 varieties into four stages (II-V) of Schneider's (2003, 2007) Dynamic Model. Applying methodology employed by Hoffmann et al. (2019, 2020), I used covarying-collexeme analysis (Stefanowitsch and Gries 2005) to test for cross-clausal associations. As it turns out, across all varieties there is statistically significant interdependence between C1 and C2 regarding multiple variables, including lexical fillers, filler types and deletion/truncation patterns. Based on these findings, I argue that previously proposed maximally abstract rules and templates are insufficient for modeling the English CC construction. Rather, my results indicate, across varieties of English, the existence of elaborate networks of so-called meso-constructions of varying degrees of abstractness. In essence, this replicates and thus corroborates Hoffmann et al.'s corpus study of the standard varieties, where comparable

meso-constructions could be detected in the corpus data (2019, 2020). I conclude that my data is best modeled as a language network consisting of interconnected constructions that is, in the words of Traugott and Trousdale, "baroque, involving massive redundancy and vastly rich detail" (2013: 53). Being the first study to examine the meso-constructional network of the CC in varieties other than British and American English, my study not only demonstrates that Hoffmann et al.'s methodology can be successfully replicated. It also makes an important contribution to constructionist approaches' "aspirations towards universal applicability" (Fried 2017: 249) by exploring non-standard varieties of a language.

References

Borsley, Robert D. 2004. An Approach to English Comparative Correlatives. In Stefan Müller (ed.), *Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar, Center for Computational Linguistics, Katholieke Universiteit Leuven*, 70–92. Stanford, CA: CSLI Publications.

Bybee, Joan L. 2012. Domain-general processes as the basis for grammar. In Maggie Tallerman & Kathleen R. Gibson (eds.), *The Oxford Handbook of Language Evolution*, 528–536. Oxford: Oxford UP.

Croft, William. 2013. Radical construction grammar. In *The Oxford Handbook of Construction Grammar*, 211–232. Oxford: Oxford UP.

Culicover, Peter W. & Ray Jackendoff. 1999. The View from the Periphery: The English Comparative Correlative. *Linguistic Inquiry* 30(4), 543–571.

Dikken, Marcel den. 2005. Comparative Correlatives Comparatively. *Linguistic Inquiry* 36(4), 497–532.

Fried, Mirjam. 2017. Construction Grammar in the Service of Slavic Linguistics, and Vice Versa. *Journal of Slavic Linguistics* 25(2), 241–276.

Goldberg, Adele E. 2003. Constructions: A New Theoretical Approach to Language. *TRENDS in Cognitive Sciences* 7(5), 219–224.

Hoffmann, Thomas, Thomas Brunner & Jakob Horsch. 2020. English Comparative Correlative Constructions: A Usage-based account. *Open Linguistics* 6(1), 196–215.

Hoffmann, Thomas, Jakob Horsch & Thomas Brunner. 2019. The More Data, The Better: A Usage-based Account of the English Comparative Correlative Construction. *Cognitive Linguistics* 30(1), 1–36.

Schneider, Edgar W. 2003. The Dynamics of New Englishes: From Identity Construction to Dialect Birth. *Language* 79(2), 233–281.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties of English around the World*. Cambridge: Cambridge UP.

Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying Collexemes. *Corpus Linguistics and Linguistic Theory* 1(1), 1–43.

Traugott, Elizabeth & Graeme Trousdale. 2013. *Constructionalization and Constructional Changes* (Oxford Studies in Diachronic & Historical Linguistics 6). Oxford: Oxford UP.

The grammar of causation: Pattern, construction, system

Susan Hunston (University of Birmingham)

The paper demonstrates how a descriptive corpus study may contribute to non-corpus-based theoretical positions. The corpus study is the Cobuild Pattern Grammar project (Francis et al. 1996; Hunston & Francis 2000); the theories are Construction Grammar (Goldberg 2006) and Systemic-Functional Grammar (Halliday & Matthiessen 2014). The current study is qualitative / interpretative. It works with eight semantic fields. For illustration, this paper focuses on the semantic field 'causation' and answers three research questions:

1) Which verb complementation patterns express causation?
2) What verb-argument constructions express causation?
3) How can the meaning of causation be modelled as a semantic network?

The starting point is 50 complementation patterns used to annotate verb senses in the Cobuild dictionaries of English since 1995. The patterns were identified on formal criteria alone, but Francis et al. (1996) divided the verbs annotated with each pattern into groups based on meaning. Twenty-five of the verb complementation patterns include meaning groups that mention causation (answering RQ1).

This study exploits the common ground between 'meaning group' and 'construction'. For this paper, 'construction' is a unit that matches form and meaning, irrespective of whether the meaning is derivable from the construction constituents. Only verb argument constructions are included. Starting from the meaning groups in Francis et al. (1996), I have identified, though not proven, almost 750 potential verb argument constructions. Thus, the original corpus research has been reinterpreted in terms of Construction Grammar. Each construction is used with several verbs. A total of 105 of the constructions express causation (answering RQ2). For example, the **Verb+noun+*into*+noun** pattern contributes to 6 constructions meaning 'cause someone to do something', comprising 61 observed verbs. Semantic roles associated with causation have been mapped on to each construction e.g. NP1: Cause; NP2: Affected; NP3: Result.

The 105 constructions have then been arranged into a network (the concept adopted from Systemic-Functional Linguistics) that, although subjective, makes sense of the verb resources in English used to express causation, showing similarities and distinctions (answering RQ 3). The paper will illustrate the major types of distinction proposed. The primary distinction focuses on what is caused: state; thought / emotion; or action/event. Within 'cause thought', a distinction of form is made between constructions that express a Cogniser (e.g. 'persuade someone that') and those that do not (e.g. 'cast doubt on something'). Animacy constitutes another distinction. Within 'cause action', for example, constructions with the pattern **Verb+noun+infinitive** are divided between those with an animate entity as the cause (e.g. 'She made him cry') and those with an inanimate entity as cause (e.g. 'The extreme cold made the engine misfire'). In other cases, a distinction between 'congruent' and 'metaphoric' is used (cf. Halliday & Martin 1993). For example, the 'cause state' constructions are divided between those where cause and result are explicit in the clause (e.g. 'She made him sad') and those where the result is implicit (e.g. 'She rid him of his insecurities'). Thus, the original corpus research has been reinterpreted as a resource for modelling the lexical end of the lexicogrammar continuum.

References

Francis G., Hunston S. and Manning E. 1996. *Collins Cobuild Grammar Patterns 1: Verbs*. London: HarperCollins.

Goldberg A. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Halliday M.A.K. and Martin J.R. 1993. *Writing Science: Literacy and Discursive Power*. London: Falmer Press.

Halliday M.A.K. and Matthiessen C. 2014. *Introduction to Functional Grammar*. 4th edition. London: Routledge.

Hunston S and Francis G. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins.

Tracking Irish English habitual *Do*: *Do* as a marker of habitual aspect in the 1641 depositions

Seamus Johnston[1], Zeltia Blanco-Suárez[2] and Teresa Fanego[2]

([1]Reitaku University, [2]University of Santiago de Compostela)

The marking of habitual aspect is one of the defining features of Irish English (IrE). A habitual marker "describe[s] a situation which […] is viewed not as an incidental property of the moment but as a characteristic feature of a whole period" (Comrie 1976:27–28). Examples (1)–(2) below illustrate two of several verbal markers employed with habitual function in IrE; respectively unemphatic *do*+lexical verb and 'invariant' *be* (see Kallen 1989, 2013: 85–106; Filppula 1999: 130–150; Hickey 2007: 213–237; Ronan 2011):

(1) More and Browns ownded it [a place in Co. Wicklow]. Guinnesses **did own** it one time. They sold it to Lord Avonmore. (Filppula 1999:134)
(2) A lot of them **be** interested in football matches. (Filppula 1999:136)

This presentation is concerned with instances of *do*+lexical verb in the 1641 Depositions, since it antedates all other IrE habitual markers, with instances being found from the early eighteenth century (Bliss 1979: 292–294; Filppula 1999: 138–139; Hickey 2007: 219–220). The 1641 Depositions are a compilation of witness testimonies recorded after a rebellion in Ireland in 1641.

Two other motivations for our choice of topic are, first, that there is consensus in the field that IrE habitual *do* stems from Early Modern British English affirmative declarative clauses featuring the verb *do* as an unemphatic, colourless auxiliary (so called 'periphrastic *do*'), a usage now obsolete. The hypothesis is that periphrastic *do* formed part of the English input to Ireland carried by speakers from the south west, and was eventually co-opted for habitual use (Filppula 1999: 136–144; Hickey 2007: 220–222). Working on this assumption, it is advantageous to have at one's disposal studies on BrE periphrastic *do*, such as Rissanen (1991), Wischer (2008), Budts/Petré (2020: 333–344), or Budts (2022), among others. Rissanen's quantitative analysis, based on the *Helsinki Corpus of English Texts* (1500–1710), concluded that he could not find "any aspectual tendency in the use of periphrastic *do* in EModE" (1991: 323).

The second motivation for the focus on habitual *do* pertains to the textual record of Early Modern IrE. Although scant, the availability since 2010 of a digital edition of the 1641 Depositions has opened new possibilities for linguistic research. For this presentation, we are employing a subset of the Depositions totalling 300,000 words and consisting of testimonies of deponents residing or based in counties Clare, Dublin, Kerry and Wicklow (mirroring the composition of Filppula's [1999] twentieth-century corpus). Our findings reveal a considerable number of habitual uses in both the present and past tenses, as (3) below. The use of *do*+lexical verb as a habitual marker at such an early stage challenges earlier accounts and suggests that the exact functions of periphrastic *do* in the input variety, British English, need to be investigated further, so as to illuminate the genesis of IrE habitual *do*.

(3) This examinett being a dweller herin Dublin […] and having some stock of cattle and other goods and monyes due to him in the Com of Roscomman where this examynett **did dwell** before his coming he there to Dublin vppon the begining of this Rebellious Inserection […] (Information of George Davys; Dublin, undated; 1641 Depositions, MS 830, fol 010r)

References

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

Bliss, Alan. 1979. *Spoken English in Ireland 1600–1740. Twenty-seven Representative Texts*. Dublin: The Dolmen Press.
Budts, Sara. 2022. A connectionist approach to analogy. On the modal meaning of periphrastic DO in Early Modern English. *Corpus Linguistics and Linguistic Theory* 18(2): 337–364.
Budts, Sara & Peter Petré. 2020. Putting connections centre stage in diachronic Construction Grammar. In Lotte Sommerer & Elena Smirnova (eds.), *Nodes and Networks in Diachronic Construction Grammar.* Amsterdam: John Benjamins, 317–351.
Comrie, Bernard. 1976. *Aspect.* Cambridge: Cambridge University Press.
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. 1989. Tense and aspect categories in Irish English. *English World-Wide* 10(1): 1–39.
Kallen, Jeffrey L. 2013. *Irish English Volume 2: The Republic of Ireland.* Berlin: Mouton de Gruyter.
Rissanen, Matti. 1991. Spoken language and the history of *do*-periphrasis. In Dieter Kastovsky (ed.), *Historical English Syntax.* Berlin: Mouton de Gruyter, 321–342.
Ronan, Patricia. 2011. Irish English habitual 'do be': More on origins and use. *Groninger Arbeiten zur germanistischen Linguistik* 53(2): 105–118.
Wischer, Ilse. 2008. What makes a syntactic change stop? On the decline of periphrastic *do* in Early Modern English affirmative declarative clauses. *Studia Anglica Posnaniensia* 44: 139–154.

Self-initiated L2 English activities and their effects on lexical complexity in student writing

Henrik Kaatari[1], Tove Larsson[2], Ying Wang[3], Pia Sundqvist[4] and Taehyeong Kim[2]

([1]University of Gävle, [2]Northern Arizona University, [3]Karlstad University, [4]University of Oslo)

Frequent engagement in extramural English (EE) activities (i.e., English-language activities that students engage in outside of the classroom) has been shown to positively influence L2 students' receptive and productive skills (e.g., Sundqvist, 2009, 2019; Sylvén & Sundqvist, 2012). There are also indications in previous studies that the *type* of EE input students receive affects their production. For example, Kaatari et al. (2023) found that written input (reading) had a positive impact exclusively on students' noun phrase complexity, whereas spoken input (conversation and watching) was associated exclusively with more diverse vocabulary. Written input may thus be expected to result in more frequent use of features commonly associated with academic writing, while spoken input may be expected to result in a broader, though not necessarily more advanced or sophisticated, vocabulary.

The present study starts off where Kaatari et al. (2023) left off by systematically testing the role of the *type* of input students receive through EE activities focusing specifically on lexical complexity. As lexical complexity has been shown to be correlated with writing quality (Kyle & Crossley, 2016), investigating the relationship between EE activities and lexical complexity seems like a fruitful next step toward increasing our understanding of the specific role that different EE activities play across student levels. In order to cover more of the construct of lexical complexity, we extend Kaatari et al.'s (2023) study of lexical diversity to also include lexical sophistication and by including a wider range of student levels. We look at junior and senior high school student writing in L2 English from the Swedish Learner English Corpus (SLEC; Kaatari et al., forthc.). SLEC contains information about how many hours per week students engage in five EE activities: reading, watching, conversation, social media, gaming. We use three types of lexical sophistication measures that have been shown in the psycholinguistics literature to have

high validity: contextual distinctiveness (Nelson et al., 1998), concreteness (Brysbaert et al., 2014), and age of exposure (Dascalu et al., 2016). We also use one measure of lexical diversity (moving average type-token ratio; Covington & McFall, 2010). Specifically, we ask the following research questions that also serve as our hypotheses:

1) Does frequent engagement with spoken conversation (conversation and watching) result in a higher degree of linguistic diversity than other types of EE exposure?
2) Does frequent engagement with longer written input (reading) result in a higher degree of linguistic sophistication, than other types of EE?
3) Does lexical complexity improve steadily across student levels?

To test these specific hypotheses, we use Structural Equation Modeling (SEM; see Larsson et al., 2021). Competing measured variable path analysis models were fitted, systematically looking at the hypothesized effects of the different EE activities on the four complexity measures. The best-fitting model (χ^2: 0.14, df: 20, CFI: 0.99, RMSEA: 0.033[0.00–0.064], SRMR: 0.067) confirmed all three of our hypotheses. It thus seems crucial to avoid grouping EE activities together into a single category, but instead consider what type of input students are exposed to. We also discuss implications for teachers.

References

Brysbaert, M., Warriner, A.B. & Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46, 904–911.

Dascalu, M., McNamara, D.S., Crossley, S. & Trausan-Matu, S. 2016. Age of Exposure: A model of word learning. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2928–2934.

Kaatari, H., Wang, Y. & Larsson, T. Forthcoming. Introducing the Swedish Learner English Corpus: A corpus that enables investigations of the impact of extramural activities on L2 writing. *Corpora* 19(1).

Kaatari, H., Larsson, T., Wang, Y., Acikara-Eickhoff, S. & Sundqvist, P. 2023. Exploring the effects of target-language extramural activities on students' written production. *Journal of Second Language Writing 62*, 101062.

Kyle, K., & Crossley, S. 2016. The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing* 34, 12–24.

Larsson, T., Plonsky, L. & Hancock, G. 2021. On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory* 17(3), 683–714.

Nelson, D.L., McEvoy, C.L. & Schreiber, T.A. 1998. The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/

Sundqvist, P. 2009. *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary*. Karlstad University Studies, 2009:55.

Sundqvist, P. 2019. Commercial-off-the-shelf games in the digital wild and L2 learner vocabulary. *Language Learning & Technology* 23(1), 87–113.

Sundqvist, P. & Wikström, P. 2015. Out-of-school digital gameplay and in-school L2 English vocabulary outcomes. *System* 51, 65–76.

Sylvén, L.K. & Sundqvist, P. 2012. Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL* 24(3), 302–321.

Clause-final *so*: Emergence and function of a new discourse marker use

Gunther Kaltenböck (University of Graz)

While the discourse marker so has received considerable attention in the past decades (e.g. Raymond 2004, Bolden 2009, Denison 2020), its use as a clause/utterance-final particle has been noted, if at all, only in passing (e.g. cursory remarks in Schiffrin 1987, Cheshire & Williams 2002). This paper focuses on precisely this use, as illustrated by the example in (1), which is shown to have substantially increased in recent spoken (American) English.

(1) AL-ROKER: And how are you celebrating your anniversary?
CRAIG-MELVIN: We're going to go down to D.C. for the Nationals game.
SHEINELLE-JONES: Oh, that's fun.
HODA-KOTB: Cool.
CRAIG-MELVIN: Going to hopefully go watch some history being made in Washington, see the Nats go to the World Series. She has covered the Nats for a number of years, **so**.
SHEINELLE-JONES: This is a special place for you guys.
AL-ROKER: Yeah
(COCA:2019:Spoken)

Drawing on a number of corpora of mainly American English (*Corpus of Contemporary American English* (COCA), *Corpus of Historical American English* (COHA), *Fisher Corpus*; as well as the *British National Corpus*), the paper addresses the following research questions: (i) What are the discourse functions of this recent innovation?, (ii) what is its typical prosodic realization?, and (iii) how can we explain its recent emergence and development? The approach is thus corpus-based (both quantitative and qualitative) and the overall framework usage-based, discourse-analytic, and interactional.

The discourse function is identified as being both interpersonal and textual. On the interpersonal level, clause-final *so* signals that the speaker relinquishes their turn and tries to elicit a response from the interlocutor. The overwhelming majority of clause-final *so* (97% of the total of 979 instances in COCA Spoken) are in fact turn-final, involving full speaker change (rather than just backchannelling). On the textual level, it ties the host clause proposition to the preceding text by providing a relevance link, which is frequently that of a warrant or an explanation (e.g. in (1), where the underlined host clause explains why they are 'going down to D.C. for the Nationals game for their anniversary').

The prosody of *so*, which will be illustrated by PRAAT pictures, typically exhibits a falling (or, less frequently, level) contour and may be prosodically integrated with the host clause. In terms of its diachronic development, clause-final *so* shows a significant increase in frequency particularly since the early 2000s, from 0.8 occurrences per million words in the period 1990-94 to 14.1 in 2015-19 in COCA Spoken (this corresponds to a significant rise also if measured against the baseline of all types of final *so*). It is argued that the emergence of clause-final *so* can be explained in terms of cooptation (Heine et al. 2015) of result subordinator uses of *so* and subsequent grammaticalization to different degrees.

References

Cheshire, Jenny & Ann Williams. 2002. Information structure in male and female adolescent talk. *Journal of English Linguistics* 30(2): 217-238.

Denison, David. 2020. Explaining explanatory so. In Ewa Jonsson & Tove Larsson (eds.). *Voices past and present: Studies of involved, speech-related and spoken texts. In honor of Merja Kytö*, 207–25. Amsterdam: John Benjamins.

Heine, Bernd, Tania Kuteva & Gunther Kaltenböck. 2015. On the evolution of final particles. In Sylvie Hancil (ed.). *Final particles*. Amsterdam: Benjamins, 111-140.
Raymond, Geoffrey. 2004. Prompting action: The stand-alone 'so' in ordinary conversation. *Research on Language and Social Interaction* 37(2): 185-218,
Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

English *Why not?* fragment questions: A corpus-based perspective

Okgi Kim and Jong-Bok Kim (Kyung Hee University)

The *why not*-question can be used as either an anaphoric information-seeking question or a rhetorical question, as illustrated by the attested data (Merchant 2006, Kramer & Rawlins 2009, Hofmann 2018, Stockwell 2022):

(1) a. I can't sleep. Why not? (COCA 1997 TV)
b. Shall we go in? Why not? (COCA 2016 MOV)

As illustrated by (1) and previous literature (Hofmann 2018), the anaphoric type requires a negative antecedent, whereas the rhetorical one can have a modalized antecedent.

We have performed a corpus investigation of the construction using COCA (Corpus of Contemporary American English). We identified 500 randomly selected tokens and analyzed these with three key variables: anaphoric vs. rhetorical reading, negativity of the antecedent, and islandhood. We found 276 tokens of anaphoric uses and 224 tokens of rhetorical uses. The construction is predominantly used in informal contexts (spoken 74%, written 26%).

The negativity variable shows us that the anaphoric type only has a negative antecedent while, as in (2), the rhetorical one can have not only a positive antecedent (202 tokens) but also a negative one (22 tokens).

(2) a. Sure I do it, man. Why not? (COCA 1992 MAG)
b. Let's try it. Why not? (COCA 2014 TV)
c. Are you okay? Sure, why not? (COCA 1997 MOV)
d. Men don't wear makeup. But why not? People in the 18th century wore makeup. (COCA 2016 MOV)

To identify the sources of the negativity in anaphoric uses, we further identified the types of negative expressions:

Table 1. Negation types of anaphoric *why not*'s antecedent

sentential negation	*not*	225 (81.5%)
	never	19 (6.9%)
	no	8 (2.9%)
	nothing	5 (1.8%)
constituent negation	*not*	17 (6.2%)
lexical negation	*unable*	2 (0.7%)

The dominant source is sentential negation expressed by *not*, but other types like negative adverbs, negative quantifiers, constituent negation, and lexical negation can also contribute to the negativity of the antecedent clause, as in (3).

(3) a. I never show my work. Why not? (COCA 2006 FIC)
b. No more samples today. Why not? (COCA 2019 TV)
c. I got nothin' to lose. Why not? (COCA 1995 MOV)
d. I was thinking of not going. Why not? (COCA 2018 MOV)
e. Computer, bring the subspace transmitter on line. Unable to comply. Why not? (COCA 2001 TV)

Moreoever, the corpus search yields 8 examples where the negative source is within an island:

(4) a. Very desirable woman who'll never go out with me. Why not? (COCA 1995 TV)
b. Something that a gentlewoman must not do to a gentleman. Why not? (COCA 2004 MOV)

Our corpus data argue against a purely syntactic analysis where the construction has a sentential source and is derived via a deletion operation, as exemplified by (5) (Hoffmann 2018, Stockwell 2022).

(5) $[_{CP}$ Why $[_{\Sigma P}$ not$_{[uNeg]}$ $[_{\text{~~TP~~}}$ ~~you can't$_{[iNeg]}$ sleep~~$]]]$

Departing from this syntax-based direction, we suggest that the anaphoric uses of the construction are directly generated by referring to a positive antecedent evoked from a negative presupposition. This non-elliptical discourse-based analysis could also account for rhetorical uses of the construction, interacting with speech acts.

References

Hofmann, Lisa. 2018. *Why not?* – polarity ellipsis in *why*-questions. Ms., University of California, Santa Cruz.

Kramer, Ruth & Kyle Rawlins. 2009. Polarity particles: an ellipsis account. In Anisa Schardl, Martin Walkow & Muhammad Abdurrahman (eds.), *Proceedings of NELS,* volume 39, 479– 492.

Merchant, Jason. 2006. Why no(t)? *Style* 40(1 & 2): 20–23.

Stockwell, Richard. 2022. Elliptical *why not*. Ms., University of Oxford, Christ Church.

Newswriting in the Caribbean diaspora: Americanization and other trends in *The Panama Tribune*

Catherine Laliberté and Diana Wengler (LMU Munich, University of Regensburg)

In recent years, research in World Englishes has experienced a diachronic turn, whereby new historical corpora have emerged from previously untapped sources, such as press materials (e.g. Hackert & Wengler 2022, van Rooy & Wasserman 2014). This élan stems from the need for real-time studies in order to describe language change in lesser-documented varieties. For many such varieties, the press often constitutes the only known or accessible source of historical language. Luckily, the press is also a genre said to be particularly open to innovation (cf. Hundt & Mair 1999). In the context of World Englishes, such historical newspaper corpora enable the study of trends in the norm orientation of (post)colonial communities, putting evolutionary models of postcolonial Englishes to the test (cf. Schneider 2007) and refining our understanding of phenomena such as Americanization. Caribbean newswriting, in particular, has been the subject of a growing body of research on norm orientation (e.g. Deuber et al. 2022, Hackert 2015, Hackert & Deuber 2015).

The present contribution introduces two small special-purpose corpora compiled from a historical newspaper published by – and for – the Afro-Caribbean diaspora of Panama. This community emerged when around 200,000 people migrated from the former British West Indies to constitute the main labor force in the construction of the Panama Canal between 1904 and 1914. As "the voice of the Afro-Antillean community on the Isthmus of Panama" (Guerrón Montero 2020: 40), *The Panama Tribune* became an important community-building and organizing force until its closure in 1972. The *Tribune* was carefully archived, which permitted the compilation of two 180,000-word corpora, for the years 1932 and 1967.

This study aims to leverage *The Panama Tribune* as a rare source of early and diasporic Caribbean English writing to complement our understanding of the creation and evolution of norms in (post)colonial contexts. Taking advantage of the time depth afforded by the corpora, the present study describes diachronic trends in the use of features that have been treated as indicative of U.S., British, and local normative orientations, such as spelling, vocabulary, BE passives, the use of the subjunctive and pseudotitles, among others (cf. Deuber et al. 2022). As an important trans-Caribbean news outlet, the normative orientation of the *Tribune* is expected to be in line with what has been observed for Caribbean newswriting, such as in the Bahamas and Trinidad and Tobago, where American norms tend to increasingly prevail, although some traditionally British aspects and local specificities may also be prominent (cf. Hackert & Deuber 2015). The Panamanian Afro-Caribbean community had ties to the pre-independence British West Indies and was compelled to devise a sense of identity on isthmian soil, straddling the U.S.-controlled Canal Zone and the Republic of Panama. This makes the conservation of British norms, radical Americanization and endonormative developments equally plausible.

References

Deuber, Dagmar, Stephanie Hackert, Eva Canan Hänsel, Alexander Laube, Mahyar Hejrani & 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. DOI: 10.1215/00031283-8791736.

Guerrón Montero, Carla. 2020. *From temporary migrants to permanent attractions: Tourism, cultural heritage, and Afro-Antillean identities in Panama*. Tuscaloosa: The University of Alabama Press.

Hackert, Stephanie. 2015. Pseudotitles in Bahamian English: A case of Americanization? *Journal of English Linguistics* 43: 143-67. DOI: 10.1177/0075424215577966.

Hackert, Stephanie & Dagmar Deuber. 2015. American influence on written Caribbean English: A diachronic analysis of newspaper reportage in the Bahamas and in Trinidad and Tobago. In Peter Collins, ed. *Grammatical Change in English World-Wide*. Amsterdam: Benjamins, 389-410. DOI: 10.1075/scl.67.16hac.

Hackert, Stephanie & Diana Wengler. 2022. Recent grammatical change in postcolonial Englishes: A real-time study of genitive variation in Caribbean and Indian newswriting. *Journal of English Linguistics* 50/1: 3-38. DOI: 10.1177/00754242211052490.

Hundt, Marianne & Christian Mair. 1999. 'Agile' and 'uptight' genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4:2, 221-242.

Schneider, Edgar W. 2007. Postcolonial English: Varieties around the world. Cambridge: Cambridge University Press.

van Rooy, Bertus & Ronel Wasserman. 2014. Do the modals of Black and White South African English converge? *Journal of English Linguistics* 42(1). 51-67. DOI: 10.1177/0075424213511463.

Stylistic variation in (fan)fiction: A stylometric analysis of original and derived fictional texts

Daniela Landert and Lea Kyveli Chrysanthopolou (University of Heidelberg)

Writers of fan fiction first and foremost try to evoke the world of a work of fiction: the characters, the locations, and the social and physical rules that structure the fictional world. At the same time, the linguistic style of the work of fiction is relevant as well. While previous research has sometimes treated fan fiction as a homogenous genre (see Girouard et al. 2013, Mattei et al. 2020), we assume that the style of writing varies across fan fictions and that for a fan fiction to be successful, the style of writing should be in line with the style of the original work. This is precisely what our analysis wishes to investigate, leading to the following research questions: Which stylistic aspects need to match the original work for a fan fiction to be successful? Are there linguistic features that can deviate from the original work without affecting the perception of fan fiction negatively? To what extent do necessary and optional stylistic features vary across different works of fiction and across different genres? And, finally, are there stylistic features that are universal to successful fictional writing?

In our study, we approach these questions from a corpus-linguistic perspective. Based on a corpus of twelve book series of original works (12.5 mio words) and their approximately 60,000 fan fictions (660 mio words), we quantify the similarities between texts based on a stylometric analysis. The original fictional texts come from the genres Fantasy, Science Fiction, and Romance. This makes it possible for us to gain insight into variation within and across genres. The fan fictions are taken from AO3 (Organisation for Transformative Works 2007), in agreement with the platform's terms and conditions and using a web-scraper by Li and Sterman (2022). The data are then processed in Python, using the Natural Language Toolkit (Bird et al. 2009). One of the features of the AO3 platform is that readers can rate texts by assigning "kudos" to them. Thus, the number of kudos a text receives provides an indication of how good the fan community considers it to be. For our analysis, we use this rating to divide the fan fictions into three separate groups: good (top 15%), intermediate (next 35%) and bad (bottom 50%). We then compare the stylistic characteristics of each fan fiction to those of the original work and calculate similarity scores. Our analysis is based on Burrow's Delta (Burrows 2002), one of the most robust measures of intertextual distance (see Neal et al. 2018: 12). We adjust the features for interpretability and focus on features that are related to literary style.

Our results show that there are various dimensions that influence which stylistic characteristics correlate with a positive rating. There are stylistic characteristics that are typical of good fan fiction writing overall, whereas other stylistic characteristics are specific to a given genre or even to a given work. The methodological contribution of our study is to determine linguistic features that are, at the same time, stylistically relevant and that can be quantitatively analysed by computational methods. For this, we draw on the fields of stylometry, corpus-based register variation and literary stylistics.

References

Bird, Steven, Ewan Klein and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Burrows, John. 2002. "Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17 (3): 267–87.

Girouard, Vanessa and Victoria Rubin. 2013. Com*parative Stylistic Fanfiction Analysis: Popular and Unpopular Fics across Eleven Fandoms*.

Li, Jingyi and Sarah Sterman. (2016) 2023. AO3Scraper. Python. https://github.com/radiolarian/AO3Scraper.

Mattei, Andrea, Dominique Brunato and Felice Dell'Orletta. 2021. The Style of a Successful Story: A Computational Study on the Fanfiction Genre. In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-It 2020 : Bologna, Italy, March 1-3, 2021*, edited by Johanna Monti, Fabio Tamburini, and Felice Dell'Orletta, 284–89. Collana Dell'Associazione Italiana Di Linguistica Computazionale. Torino: Accademia University Press. https://doi.org/10.4000/books.aaccademia.8718.

Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang and Damon Woodard. 2017. Surveying Stylometry Techniques and Applications. *ACM Computing Surveys (CSuR)* 50 (6): 1–36.

Authenticity in country music: A corpus-based perspective on styling

Anna Ledermann and Valentin Werner (University of Bamberg)

Country music, originally associated with the US South(west), has become a commercially highly successful genre both in the US and worldwide (Nielsen 2022). Country has been perceived as valuing authenticity, reflected in distinctive song themes such as (white) rural and working-class origins and lives (Fox 2004), a typical choice of instrumentation or an artist's clothing style, but importantly also in the choice of linguistic features in the lyrics (Duncan 2017).

With (white) Southern American English (SAE) serving as the "default" variety, it has been observed that artists who are not SAE speakers must use some of its features in order to be successfully perceived as authentic country performers (Davies & Myrick 2018). However, given the recent increasing diversification of the genre (see, e.g., Bates et al. 2020) and the fact that by no means all current country artists are white Americans from the South, the question arises as to whether the use of SAE features is still considered obligatory as a kind of "supralocal norm" (Gibson 2023). While previous work has largely focused on pronunciation (see, e.g., Horn 2010; Duncan 2017), the present study considers morphosyntactic features to determine whether and how they are used to index authenticity in country lyrics, as has been traced for other musical genres (Werner 2019).

To this end, it compares the lyrics of country songs by white southern, non-white southern, and white non-southern male and female artists with respect to their use of 12 morphosyntactic SAE features identified in the Yale Grammatical Diversity Project (YGDP 2023), such as *a*-prefixing, personal datives, *what all*-constructions, etc. The data used for the analysis consists of the lyrics of 600 highly successful songs featuring on the Billboard "Hot Country" year-end charts (2000–2022), supplemented by listings on the specialized website newcountrysongs.com. Lyrics were obtained from azlyrics.com, normalized for spelling variations, cleaned of metainformation, and then tagged for parts of speech using CLAWS (Garside & Smith 1997). The relevant morphosyntactic structures were retrieved from the corpus by queries in AntConc (Anthony 2023).

The basic assumption tested is that if these features are evenly distributed or less common in the lyrics of white southerners than in the lyrics of the other two groups, this would provide evidence for their function of indexing authenticity in the sense of having become enregistered for this music genre (Agha 2005). Alternatively, a growing diversity of the country genre might be reflected in the grammar of country songs if SAE features are less salient in the lyrics of non-white southern and white non-southern artists. A secondary goal of the present study is to complement previous case studies of styling, authenticity, and enregisterment in other genres like rock (Flanagan 2019), blues (Larroque 2022), folk (Watts & Morrissey 2021), or rap (Werner 2019; Gibson 2023), with a corpus-based perspective.

References
Agha, Asif. 2005. Voice, footing, enregisterment. *Journal of Linguistic Anthropology* 15(1): 38–59.
Anthony, Lawrence. 2023.*AntConc* (Version 4.2.2). Tokyo: Waseda University. https://laurenceanthony.net/software/antconc/
Davies, Catherine E. & Caroline Myrick. 2018. Performing southernness in country music. In Jeffrey Reaser, Eric Wilbanks, Karissa Wojcik & Walt Wolfram (eds.), *Language variety in the new South: Contemporary perspectives on change and variation*, 78–96. Chapel Hill: The University of North Carolina Press.
Bates, Vincent C., Jason B. Gossett & Travis Stimeling. 2020. Country music education for diverse and inclusive music classrooms. *Music Educators Journal* 107(2): 28–34.
Duncan, Daniel. 2017. Australian singer, American features: Performing authenticity in country music. *Language & Communication* 52: 31–44.
Flanagan, Paul J. 2019. "A Certain Romance": Style shifting in the language of Alex Turner in Arctic Monkeys songs 2006–2018. *Language and Literature* 28(1): 82–98.
Fox, Aaron A. 2004. *Real country: Music and language in working-class culture*. Durham: Duke University Press.
Garside, Roger & Nicolas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In Roger Garside, Geoffrey Leech & Tony McEnery (eds.), *Corpus annotation: Linguistic information from computer text corpora*, 102–121. London: Longman
Gibson, Andy. 2023. Pop song English as a supralocal norm. *Language in Society*. https://doi.org/10.1017/S0047404523000131
Horn, Alena. 2010. "I can't help it the way I talk": Dialect, language attitudes, and style shift in country music. In Patricia Donaher (ed.), *Barbarians at the gate: Studies in language attitudes*, 158–182. Newcastle: Cambridge Scholars.
Larroque, Patrice. 2022. English rhythm and blues: Where language and music come together. London: Routledge.
Nielsen. 2022. Distribution of streamed music consumption in the United States in 2021, by genre. https://www.statista.com/statistics/475667/streamed-music-consumption-genre-usa/
Watts, Richard J. & Franz Andres Morrissey. 2021. *Language, the singer and the song: The sociolinguistics of folk performance.* Cambridge: Cambridge University Press.
Werner, Valentin. 2019. Assessing hip-hop discourse: Linguistic realness and styling. *Text & Talk* 39(5): 671–698.
YGPD. 2023. Yale Grammatical Diversity Project – Phenomena: Southern American English. https://ygdp.yale.edu/phenomena-by-category?field_phenomenon_category_value_1=All&field_phenomenon_dialect_value_1=10

Ready for the sufferfest: A corpus-based study of *-fest* in the specialised register of mountaineers

Sven Leuckert (TU Dresden)

This study investigates a potential source of lexical innovation in the specialised register of English-speaking mountaineers, which I henceforth refer to as 'Mountaineering English' (MountE). The feature in focus is the morpheme *-fest* (as in *gabfest* and *gorefest*), a form that has periodically been noted as being part of colloquial American English (AmE) (e.g., Hirshberg 1981; Lazerson 1984; Green 2023). Rather intriguingly, the *Climbing Dictionary: Mountaineering Slang, Terms, Neologisms & Lingo* (Samet 2011) lists *-fest* as one of the headwords (occurring in combinations such as *crimpfest*, *pumpfest*, and *takefest*), suggesting it is a prominent form in MountE. However, so far, there is no corpus-based evidence to support its status as a feature of the specialised register. Furthermore, the morphological status of *-fest* is contested, with

publications variously describing it as combining form (e.g., Mattiello 2017) or as an affix (e.g., O'Dell 2016).

In order to learn about the forms and frequency of *-fest*, I report on the results of a corpus-based study. The data come from five online forums (or thematic 'subreddits') dedicated to mountaineering on the social media platform Reddit (amounting to ca. 70 million words), with tokens found in the Corpus of Historical American English (COHA) and the Corpus of Contemporary American English (COCA) included for comparison. The three main research questions are:

a) How frequently do forms with *-fest* occur in a corpus of subreddits dedicated to mountaineering and do the frequencies provide empirical evidence to Samet's (2011) suggestion that it is a common feature of MountE?
b) How do the frequencies of *-fest* in the subreddits compare to frequencies in COHA and COCA and do they follow the trend of increasing frequency in (American) English more generally?
c) Which spellings are preferred in forms with *-fest* (open, hyphenated, or solid, see Sanchez-Stockhammer 2018) and what are the implications of these preferences for a morphological analysis of *-fest*?

The results show that forms with *-fest* are used roughly equally frequently in COCA and the subreddits, with an increasing frequency over time in COHA. While the productivity of *-fest* is much higher in COHA and COCA than in the five subreddits, more than 50% of all *-fest* tokens in COHA and COCA represent hapax legomena. In the subreddits, only about 11% represent hapaxes, suggesting that the forms in use are recurring and, in some cases such as *jugfest* and *sufferfest*, established forms of the variety. The findings thus indicate that *-fest* has become part of the specialised register of mountaineers, likely functioning as an in-group marker. All three spelling variants occur, but the surprisingly high frequency of open spellings (as in *ice fest*) suggests that *-fest* cannot reasonably be analysed as a combining form anymore. Instead, it may have become entrenched in MountE to the extent that it can be used as a free morpheme in compounds.

References

Green, Jonathon (2023). *Green's Dictionary of Slang*. https://greensdictofslang.com/. Accessed 28 November 2032.

Hirshberg, Jeffrey (1981). Regional morphology in American English: Evidence from Dare. *American Speech* 56(1): 33-52.

Lazerson, Barbara Hunt (1984). Another *fest*-icon. *American Speech* 59(3): 261-264.

Mattiello, Elisa (2017). *Analogy in Word-Formation: A Study of English Neologisms and Occasionalisms*. Berlin: Mouton de Gruyter.

O'Dell, Felicity (2016). Creating new words: affixation in neologisms. *ELT Journal* 70(1): 94-99.

Samet, Matt (2011). *The Climbing Dictionary: Mountaineering Slang, Terms, Neologisms & Lingo*. Seattle, WA: The Mountaineers.

Sanchez-Stockhammer, Christina (2018). *English Compounds and their Spelling*. Cambridge: Cambridge University Press.

‘Bad’ Indian English: The lexicology of multilingual swearing

Sven Leuckert and Claudia Lange (TU Dresden)

‘Bad language’ (Andersson & Trudgill 1990) has been established as an informal cover term for linguistic forms and practices ranging from nonstandard language to slang (Coleman 2012) and ‘forbidden words’ (Allan & Burridge 2006) such as expletives and swearwords (e.g. Hughes 2006). Corpus-linguistic research on swearing in varieties of English has so far been restricted to Inner Circle varieties (e.g. Love 2021; Schweinberger 2018); research on Outer Circle varieties such as Indian English (IndE) has been limited partly by the lack of suitable corpora: the design features of the ICE-corpora, for example, include a focus on educated standard(ising) language and on representativeness rather than size. This paper takes Lambert’s work on IndE slang (2014) as its point of departure, updating and extending his investigation of ‘bad’ IndE with a special focus on the multilingual range of swearwords available to the contemporary Indian English speech community.

The database to be used lends itself particularly well to the study of swearing: we investigate swear words in two subreddits, i.e., thematic forums on the social media platform Reddit, with different overt political stances. While r/indianews (13,107,349 words) presents itself as politically neutral, r/indiaspeaks (79,097,344 words) is openly nationalist. These subreddits are large enough for the study of lexis (Szmrecsanyi & Rosseel 2020: 31), and they represent interactive language use including a high share of insulting language. In order to cope with the big-data nature of the two subreddits, we identified relevant swear words by combining a word-list approach based on previous literature as well as personal communication with corpus queries for swear words in specific patterns (such as *you are (such) a **). The two main research questions we investigate are:

1) Which semantic fields do the swear words on the two subreddits belong to, and are there differences between Hindi vs. English expressions as far as these semantic fields are concerned?
2) Do the political alignments of the two subreddits – r/indianews as politically neutral and r/indiaspeaks as right-leaning – potentially lead to higher frequencies of Hindi-based swear words in r/indiaspeaks (and vice versa for English in r/indianews)?

Our preliminary results confirm that swearing in both Hindi and English is surprisingly frequent in the subreddits, both of which express in their rules that users should not be abusive or offensive (although r/indiaspeaks provides more details). Further, while IndE on Reddit shares a predilection for the lemma FUCK with informal spoken British English (Love 2021: 750), the actual realization displays nativization to different degrees: either by using a Hindi expression (e.g. *madarchod* ‘motherfucker’), a calque from Hindi such as *sisterfucke*r (see also Lambert 2014: 129), or English expressions adapted to the Indian contexts such as *cowfucker*.

References

Allan, Keith & Kate Burridge (2006). *Forbidden Words: Taboo and the Censoring of Language*. Cambridge: Cambridge University Press.

Andersson, Lars-Gunnar & Peter Trudgill (1990). *Bad Language*. Oxford: Blackwell. Coleman, Julie (2012). *The Life of Slang*. Oxford: Oxford University Press.

Hughes, Geoffrey (2006). *An Encyclopedia of Swearing: The Social History of Oaths, Profanity, Foul Language, and Ethnic Slurs in the English-Speaking World*. London: Routledge.

Lambert, James (2014). Indian English slang. In Julie Coleman (ed.), *Global English Slang: Methodologies and Perspectives*. London: Routledge, 124-134.

Love, Robbie (2021). Swearing in informal spoken English: 1990s–2010s. *Text & Talk* 41(5- 6): 739-762.

Schweinberger, Martin (2018). Swearing in Irish English: A corpus-based quantitative analysis of the sociolinguistics of swearing. *Lingua* 209: 1-20.

Szmrecsanyi, Benedikt & Laura Rosseel (2020). English corpus linguistics. In Bas Aarts, April McMahon, & Lars Hinrichs (eds.), *The Handbook of English Linguistics*. Oxford: Wiley, 29-44.

The evolution of evidential adverbs viewed through Late Modern English trials: *Evident(ly)*, *apparent(ly)*, and *clear(ly)* in the Old Bailey Corpus

Diana Lewis (Aix-Marseille University)

The rise of modal sentence adverbs in English has been well documented (e.g. Swan 1988, 2008). It has been shown that in Old English and Middle English they were restricted mostly to truth intensification, and expanded to a range of epistemic and evidential meanings in the Early Modern English period (González-Alvarez 1996). In the Modern English period they have flourished: according to Wierzbicka (2006), they developed gradually to the point where "the existence of a large class of epistemic adverbs constitutes a peculiar feature of modern English" (2006: 248; see also Simon Vandenbergen and Aijmer 2007, Kemp 2018). The aim of the present study is to characterize the semantic and syntactic evolution of three evidential markers in the Old Bailey Corpus (OBC) (Huber et al. 2016) of trial transcriptions covering the period between the 1740s and the very early 1900s. The OBC samples a particular genre, providing good diachronic genre continuity, an approximation to spoken face-to-face language, and narratives of past situations where the source and the reliability of knowledge of what happened are important. The adverbs were chosen because they are among the most frequent of evidential adverbs in the corpus, have corresponding adjectival constructions and express similar types of evidentiality.

Usage of each of the markers in this data is shown to have evolved from more concrete meanings towards modal and inferential meaning. It is found (a) that to a large extent the semantics of the adjectives (*evident*, *apparent*, *clear*) and of the adverbs (*evidently*, *apparently*, *clearly*) evolve in parallel; (b) that overall the adverbs become relatively more frequent during the period; (c) that alongside more evidential usage there is considerable persistence of non-modal usage; this might also be described as semantic expansion followed by narrowing. There is also some evidence that overall use of the adverbs by 'higher' class speakers is higher than that by 'lower' class speakers, and that the relatively lesser use by lower class speakers includes proportionally more newer evidential uses.

Following the presentation of the data, the findings are looked at in the context of models of change. Innovative modal uses of English adverbs have often been described in terms of loss of lexical or propositional meaning and increased grammaticalization. In this case, no firm correlation between syntactic change and semantic change is found, and the evolution is considered in the light of a wider phenomenon of adverbialization in the recent history of English linked to discourse prominence and information compression.

References

González-Álvarez, D. 1996. Epistemic disjuncts in Early Modern English. *International Journal of Corpus Linguistics* 1(2): 219–256.

Huber, M., M. Nissel and K. Puga. 2016. *Old Bailey Corpus 2.0.* hdl:11858/00-246C-0000-0023- 8CFB-2.

Kemp, L. 2018. English evidential -ly adverbs in main clauses: A functional approach. *Open Linguistics* 4, 743-761.

Simon-Vandenbergen, A.M. and K. Aijmer. 2007. *The Semantic Field of Modal Certainty*. Berlin: Mouton de Gruyter.

Swan, T. 1988. *Sentence adverbials in English: a synchronic and diachronic investigation*. Oslo: Novus.

Swan, T. 2008. The development of sentence adverbs in English. *Studia Linguistica* 42(1): 1-17.

Wierzbicka, A. 2006. Probably: English epistemic adverbs and their cultural significance. In A. Wierzbicka *English: Meaning and Culture*, 247–296. Oxford: Oxford University Press.

The development of *-body/-one* indefinites in eighteenth-century British publishing networks

Aatu Liimatta and Tanja Säily (University of Helsinki)

The history of English indefinites is a much-studied topic. Using the *Corpus of Early English Correspondence* and its eighteenth-century *Extension*, Nevalainen & Raumolin-Brunberg (2003) and Laitinen (2018) show that the rise of the forms in *-body* and *-one* was led by women and the higher social ranks and that *-body* was the preferred variant in correspondence, with *-one* slowly taking over. By contrast, D'Arcy et al. (2013) find that *-one* dominated in the more literate genres represented by the *Penn Parsed Corpora of Historical English* and was emerging as the 'standard' variant by 1700. The stylistic differentiation between *-one* and the more 'vernacular' *-body* persists to this day, with the shift towards *-one* still ongoing and regionally varied (D'Arcy et al. 2013; Öhman et al. 2019).

The eighteenth century represents a particularly interesting period in the development of *-body* and *-one*. Laitinen (2018) finds more sociolinguistic variation in their use in this period than Nevalainen & Raumolin-Brunberg (2003) did in Early Modern English. For instance, the clergy lagged behind in the use of *-body*, possibly because the Authorized Version of the Bible preferred *-man* and *-one*. Furthermore, Säily (2018) discovers evidence for a social network effect in individuals who were consistently progressive in terms of more than one ongoing change, including the incoming indefinites. In D'Arcy et al.'s (2013) study, there was still a great deal of variation in the use of *-body* and *-one* by quantifier, so that the proportion of *somebody* and *nobody* was on the increase, whereas *anybody* and *everybody* were decreasing. The analysis of *-body* and *-one* in published texts of the period is however lacking in that the corpora used have been relatively small, and owing to lack of suitable metadata, social aspects have been largely ignored.

In this study, we will analyse the use of *-body and -one* indefinites in eighteenth-century Britain making use of *Eighteenth Century Collections Online* (ECCO), a large-scale dataset of works published in the eighteenth century, together with metadata from an augmented version of the *English Short Title Catalogue* (Tolonen et al. 2021). We will use as our starting point the set of works associated with Andrew Millar, a prominent publisher of the period (Ryan & Tolonen 2024). We will explore the role of such publisher networks, consisting of publishers and their associated authors and printers, as a type of social network in language change. The *nodes* of the network are formed by individual book trade actors, whereas the *edges*, or the links between the nodes, are based on the co-occurrence of the individuals in the publication metadata. In particular, we are interested in the role which the connections between various actors of the publishing industry may have played in helping spread or resist linguistic innovations, testing Milroy's (1987) weak-tie hypothesis.

We perform automated searches of the indefinites of interest in their various attested spellings. We then compare their observed frequencies across e.g. time periods, authors and genres to build a picture of their use and spread. Our preliminary results indicate that in works published by Millar over the century, the *-one* variant is already dominant with the quantifiers *any* and *every*, and it is gaining ground with *no* and *some*. This could reflect the overall process of standardization of *-one* becoming increasingly dominant in literate genres over the period.

References
D'Arcy, Alexandra, Bill Haddican, Hazel Richards, Sali A. Tagliamonte & Ann Taylor. 2013. Asymmetrical trajectories: The past and present of *-body*/*-one*. *Language Variation and Change* 25(3). 287–310.
Laitinen, Mikko. 2018. Indefinite pronouns with singular human reference: Recessive and ongoing. In Terttu Nevalainen, Minna Palander-Collin & Tanja Säily (eds.), *Patterns of change in 18th-century English: A sociolinguistic approach* (Advances in Historical Sociolinguistics 8), 137–158. Amsterdam: John Benjamins.
Milroy, Lesley. 1987. *Language and social networks*. 2nd edition. Oxford: Blackwell.
Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England* (Longman Linguistics Library). London: Pearson Education.
Öhman, Emily, Tanja Säily & Mikko Laitinen. 2019. Towards the inevitable demise of *everybody*? A multifactorial analysis of *-one*/*-body*/*-man* variation in indefinite pronouns in historical American English. Paper presented at the 40th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 40), Neuchâtel, Switzerland, June 2019.
Ryan, Yann Ciarán & Mikko Tolonen. 2024. The evolution of Scottish Enlightenment publishing. *The Historical Journal* 67(2). 223–255.
Säily, Tanja. 2018. Conservative and progressive individuals. In Terttu Nevalainen, Minna Palander-Collin & Tanja Säily (eds.), *Patterns of change in 18th-century English: A sociolinguistic approach* (Advances in Historical Sociolinguistics 8), 235–242. Amsterdam: John Benjamins.
Tolonen, Mikko, Eetu Mäkelä, Ali Ijaz & Leo Lahti. 2021. *Corpus linguistics and Eighteenth-Century Collections Online* (ECCO). *Research in Corpus Linguistics* 9(1). 19–34.

Emerging evidential parentheticals in contemporary American English: Exploring COCA

María José López-Couso and Belén Méndez-Naya (University of Santiago de Compostela)

Languages deploy various methods to specify the nature of the evidence for a given statement, i.e. whether the information has been seen, heard, reported, inferred, etc. However, while some 25% of the world's languages have evidentiality as an obligatory grammatical category, others, like English, make use of "evidential strategies" (Aikhenvald 2004), including lexical devices (e.g. modal adverbs) as well as other means of expression showing different degrees of grammaticalization, such as parentheticals (e.g. *it seems*) and modals (e.g. *should*) (Chafe 1986).

This paper is concerned with emerging parenthetical structures of the type exemplified in (1)-(3), which have so far not attracted scholarly attention.

(1) When it comes to New Year's resolutions, shedding debt runs neck and neck with shedding pounds, **surveys show**. (COCA, 2007, NEWS)
(2) Psychotherapy for fear**, the research suggests,** should be coupled with healthy sleep. (COCA, 2012, BLOG)
(3) After a hectic day - when the lure of the drive-thru is most magnetic - your metabolism tanks**, says a new study.** (COCA, 2014, MAG)

Formally, in contrast with paradigmatic parenthetical clauses (those with a first- or second-person subject, such as *I think* or *you know*), the clausal parentheticals under study here feature a third-person subject with a noun denoting an examination or an investigation followed by a VP. As regards their semantico-pragmatic content, these parentheticals allow speakers/writers to present the information in the proposition as hearsay, thus avoiding responsibility for the statement, while attributing it to a reliable source.

Using mainly data from the Corpus of Contemporary American English (COCA; Davies 2008-) and focusing on evidential parentheticals with the nouns *study*, *research*, *survey*, *report*, *analysis*, and *evidence*, this paper examines: (i) the parenthetical patterns attested in the data (non-inverted (1)-(2) vs. inverted (3)); (ii) the various realizations of these patterns, regarding the complexity of the subject NP (e.g. presence/absence of determiners and modifiers) and TAM marking in the VP; (iii) the predicate-types occurring in these parentheticals; (iv) the position of the parentheticals in the sentence in relation to their host clauses; (v) their diachronic distribution; and (vi) their association with particular text-types.

Preliminary results from a pilot survey of evidentials with the noun *study* reveal a clear preference for the non-inverted pattern (over 95% of the total) as well as for final position (ca. 85%). As regards the verb, *study*-parentheticals in COCA are associated with three predicate-types, all conveying evidential meaning: utterance (e.g. *say*), demonstration (e.g. *show*), and (acquisition of) knowledge (e.g. *find*). Parentheticals with *study* are closely related to the written language, particularly to the press category (popular magazines and newspapers; over 70% of the relevant instances). Moreover, although they exhibit a certain degree of variability (in terms of the presence/absence of determiners and of modifiers in the NP and of TAM marking in the VP), the evidence also suggests that the sequence *study finds* (bare noun + present tense VP) comes close to a "formulaic thetical" (Kaltenböck et al. 2011) which, as the COCA data show, has become a staple device in journalese.

References

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford: Oxford University Press.

Chafe, Wallace. 1986. Evidentiality in English conversation and academic writing. In *Evidentiality: The Linguistic Coding of Epistemology*, Wallace Chafe & Johanna Nichols (eds.), 261-272. Norwood, NY: Ablex.

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

Kaltenböck, Gunther, Bernd Heine & Tania Kuteva. 2011. On thetical grammar. *Studies in Language* 35 (4): 848-893.

Connectors in novice and international expert EAP: BA vs. MA vs. published writing

Tomáš Mach (Charles University)

It is universally agreed that a good piece of writing has to be coherent and cohesive. To compose a text that is easy to navigate, writers employ a variety of different strategies and metadiscursive features including linkers/connectors/connectives/discourse markers, which are a frequent occurrence in (English) academic writing (Peacock, 2010). The extent to which their judicious and accurate use contributes to the overall writing quality is, however, disputed (Shea, 2009; Yang & Sun, 2012). As with many other language features, disparities in the use of connectors have been found; they are reported to be either overused, underused, or misused by novice writers.

In most studies on the subject, expertise rather than nativeness seems to be a common denominator influencing how they are deployed. Specifically, novice writers – both native and non-native – have a tendency to rely on a limited set of connectors, and to simultaneously overuse these (Lei, 2012; Shaw, 2009). These instances of overreliance have been quantified for instance by Appel & Szeib (2018) whose analysis revealed that the ten most frequent linkers in

the analysed essays amount to approximately 60-70% of all linkers utilised. This effect has been observed across different learner L1s including natives (Bolton, Nelson, & Hung, 2002), although the extent to which linkers are overused as well as to what linkers in particular this concerns varies (Appel & Szeib, 2018; Leńko-Szymańska, 2008).

Being a part of a larger project investigating L1 Czech novice academic English, this study on connectives is based on a corpus of student and expert writing (BA and MA theses, and published articles respectively) comprising 19 million words. The aims of the study are twofold; first, using a much larger sample, it seeks to examine the results of previous research on linkers in Czech EAP student writing (Povolná, 2012; Vogel, 2008), which mostly concurs with the presented international findings. Second, it aims to explore differences between BA and MA with expert published writing as a benchmark. The design of the corpus also makes it possible to track these differences down to the individual level as two thirds of theses in the corpus are BA-MA pairs produced by the same student. The final grades were recorded as well, which allows for inferences about the effect of linkers on the perceived writing success. With this in mind, the following research questions have been formulated:

1) How do L1 Czech BA and MA theses differ from expert writing in the use of connectors?
2) What is the relationship between assessment and the frequency of connectors?

In line with Liu (2008), the analysis draws on finite lists and systems of categorization from previous research, and rather than pursuing a native-non-native comparison, three levels of expertise (BA, MA, and expert) were opted for instead. The results seem to corroborate earlier findings that learners tend to use more connectors overall compared to expert writers. Minimal differences have been found between the two learner levels.

References

Appel, R. & Szeib, A. (2018). Linking adverbials in L2 English academic writing: L1-related differences. *System* 78, pp. 115–129. doi:10.1016/j.system.2018.08.008

Bolton, K., Nelson, G. & Hung, J. (2002). A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics* 7(2), 165–182. doi:10.1075/ijcl.7.2.02bol

Lei, L. (2012). Linking adverbials in academic writing on applied linguistics by Chinese doctoral students. *Journal of English for Academic Purposes* 11(3), 267–275. doi:10.1016/j.jeap.2012.05.003

Leńko-Szymańska, A. (2008). Non-native or non-expert? The use of connectors in native and foreign language learners' texts. *Acquisition et interaction en langue étrangère* 27, 91–108. doi:10.4000/aile.4213

Liu, D. (2008). Linking adverbials: An across-register corpus study and its implications. International *Journal of Corpus Linguistics* 13(4), 491–518. doi:10.1075/ijcl.13.4.05liu

Peacock, M. (2010). Linking adverbials in research articles across eight disciplines. *Ibérica* 20, 9–34.

Povolná, R. (2012). Causal and contrastive discourse markers in novice academic writing. *Brno Studies in English* 38(2), 131–148. doi:10.5817/BSE2012-2-8

Shaw, P. (2009). Linking adverbials in student and professional writing in literacy studies: What makes writing mature. In M. Charles, D. Pecorari & S. Hunston (eds.), *Academic Writing: At the Interface of Corpus and Discourse* (pp. 215–235). London: Continuum.

Shea, M. (2009). *A Corpus-Based Study of Adverbial Connectors in Learner Text*. SU Working Papers in SLS 2009, 1(1), 1–13.

Vogel, R. (2008). Sentence linkers in essays and papers by native vs. non-native writers. *Discourse and Interaction* 1(2), 119–126.

Yang, W. & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education* 23(1), 31–48. doi:10.1016/j.linged.2011.09.004

New OED entries from Nigeria and the Caribbean: Testing the selection criteria through corpora and ChatGPT

Christian Mair (University of Freiburg)

In preparation for the 2028 centenary edition, the editors of the OED are pursuing a laudable decolonisation policy by systematically extending coverage of postcolonial Englishes. Recent major updates featuring Nigerian English and Caribbean English words and senses (January 2020 and September 2022, respectively) have resulted in more than a hundred new entries, numerous additional sub-entries and new senses in existing entries. More of the new entries originate from the Caribbean than from Nigeria. For some items, there is a significant time lapse between first attestation and inclusion in the dictionary. After eliminating technical terms (e.g. from botany and linguistics), I have created balanced samples of 25 words/senses each from the Caribbean and Nigeria to check their distribution in standard reference corpora of postcolonial Englishes (ICE, GloWbE, NOW). The following trends emerge from the analysis:

1) The data from the OED and the corpora are broadly coherent, but individual words may differ drastically in their frequency of use and dispersion across registers and text-types.
2) On the basis of first attestations, both Caribbean and Nigerian items can usefully be separated into three diachronic layers: (i) a pre-1960 colonial phase, (ii) a 1960-1990 phase of postcolonial emancipation, and (iii) a globalisation phase extending from ca. 1990 to the present.
3) Many phase (iii) entries originate from Caribbean creoles and Nigerian Pidgin, sometimes mediated through local varieties of Standard English.
4) A significant part of phase (iii) entries is rapidly taken up in other varieties of English, which makes them postcolonial internationalisms.

In contrast to the traditional *OED-as-corpus* paradigm (Hoffmann 2005), the present study can be described as *OED-with-corpus*. In a second analytical round, and following Torrent et al. (2023), I use ChatGPT as a "copilot for linguists." While Nigerian and Caribbean localisms are generally difficult to prompt, this is different for many postcolonial internationalisms, which is due to their presence in diasporic identity politics, popular culture and the music and entertainment industries. The resulting selective and skewed presence of the New Englishes in Large Language Models is a challenge for World Englishes theory, as it represents a standardisation paradox. Caribbean creoles and Nigerian Pidgin are still stigmatised to some extent in vernacular usage, but have attained high visibility in globally relevant advanced language technologies.

References

Hoffmann, Sebastian. 2005. Using the OED Quotations Database as a Corpus – a Linguistic Appraisal. *ICAME Journal* 28: 17–30.

Torrent, Tiago, Thomas Hoffmann, Arthur Lorenzi Almeida and Mark Turner. 2023. *Copilots for Linguists: AI, Constructions and Frames*. Cambridge: CUP.

Willinsky, John. 1994. *Empire of Words: The Reign of the OED*. Princeton, NJ: Princeton UP.

Presentation of direct speech in crosswriters' fiction for children and for adults

Markéta Malá (Charles University)

The increase in popularity of crossover literature since the turn of the century (Beckett 2009), seems to have been accompanied by an increase in the interest in similarities and differences between fiction for children and for adults, and "[works] of crosswriters, authors who write for both readerships in different works, are an excellent source of this research" (Haverals et al. 2022: 62). The present study examines the works of four British crosswriters to explore the impact of the intended reader on the one hand, and the authorial style on the other on the presentation of direct speech in fiction.

As suggested by Stockwell and Mahlberg (2015: 130), "the relationship that readers develop with fictional characters is a main motivating factor in reading literature at all." The construction of characters relies, to a large extent, on direct speech, and the readers' interpretations of characters' speech has been shown to be "greatly influenced by the use of the speech verbs that introduce the characters' words" (Ruano San Segundo 2016: 114). These verbs "can contribute to further fleshing out a character" (ibid.), triggering information about personality (Culpeper: 215); "[the] dramatization of characters' voices is an important part of the effects of vividness, immediacy and involvement of fictional narratives" (Semino and Short 2004: 92). At the same time, "[by] examining the verbs that gloss a represented saying […] we can detect the narrator's stance towards what is reported" (Caldas-Coulthard 1988: 6, in Ruano San Segundo 2016: 115).

Methodologically, the study combines a corpus-assisted quantitative approach with a qualitative analysis of text samples. The analysis draws on texts of 5 books for children by R. Dahl, 3 by M. Paver, 2 by J.K. Rowling, 2 by S. Rushdie (total 717 thousand words), and the same numbers of novels for adults by the same authors (1.6 million words). The corpora were installed in CQPweb (Hardie 2012). The results confirm that in books for children, the proportion of direct speech is significantly higher than in adult fiction (cf. Anderson 1984: 56). Reporting verbs were identified as collocates of inverted commas, and classified using Caldas-Coulthard's (1994) taxonomy. When writing for children, the writers were found to rely most heavily on 'descriptive verbs', which refer to vocal effects and voice quality (e.g. *hissed*, *mumbled*), highlighting the importance of sound in children's literature. In their fiction for adults, 'speech reporting' metapropositional verbs explicitly indicating the intended illocutionary force (e.g. *agreed*, *accused*) were dominant. The results also underline the role of body language in presenting direct speech (cf. Korte 1997, Čermáková and Malá 2021). At the same time, the writers were found to differ in the extent of their presence in the text (e.g. in the use of glossing phrases with reporting verbs; e.g. *said Snape icily*), and the diversity of reporting verbs they employ.

The results accentuate the role of 'speech verbs' in developing the readers' relationship with characters, facilitating this important meaning-making process in fiction reading especially for 'novice readers' (Nikolajeva 2014).

References

Anderson, C.C. (1984) *Style in Children's Literature: A Comparison of Passages from Books for Adults and for Children*. Open Access Dissertations. University of Rhode Island.

Beckett, S.L. (2009) *Crossover Fiction: Global and Historical Perspectives*. New York: Garland.

Caldas-Coulthard C.R. (1988) *Reporting Interaction in Narrative: A Study of Speech Presentation in Written Discourse*. Unpublished PhD Thesis, University of Birmingham, UK.

Caldas-Coulthard, C.R. (1994) On reporting reporting: the representation of speech in factual and factional narratives. In M. Coulthard (ed.) *Advances in Written Text Analysis*. London and New York: Routledge. 295-308.

Čermáková, A. and M. Malá (2021) Eyes and speech in English, Finnish and Czech children's literature. In A. Čermáková, S. Oksefjell Ebeling, M. Levin and J. Ström Herold (eds) *Crossing the Borders: Analysing Complex Contrastive Data*, *BeLLS* 11(1): 185–208.

Hardie, A. (2012) CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380–409.
Haverals, W., L. Geybels and V. Joosen (2022). A style for every age: A stylometric inquiry into crosswriters for children, adolescents and adults. *Language and Literature* 31(1): 62-84.
Korte, B. (1997) *Body Language in Literature*. University of Toronto Press.
Nikolajeva, M. (2014) *Reading for Learning. Cognitive Approaches to Children's Literature*. Amsterdam: John Benjamins.
Ruano San Segundo, P. (2016) A corpus-stylistic approach to Dickens' use of speech verbs: Beyond mere reporting. *Language and Literature* 25(2): 113-129.
Semino, E. and M. Short (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.
Stockwell, P. and M. Mahlberg (2015) Mind-modelling with corpus stylistics in David Copperfield. *Language and Literature* 24(2): 129-147.

'Connective profiles' of the five text types across the history of English: Exploring orality and literacy

Imogen Marcus (Edge Hill University)

Many studies of clausal connectives in the history of English have focused on the development of individual connective items, or on the inventory of items available during a particular time period. They also do not tend to include twenty-first century English data in their analyses. However, corpus-based methodologies which take a broader view, such as the connective profiling approach employed in this study, can be illuminating. Creating a connective profile of a text type involves assessing the frequency and distribution of clause-connecting coordinators and subordinators within it (cf Kohnen 2007). A profile can help the researcher to empirically assess the proximity of an individual text type to the spoken or written mode, as well as what their diachronic development might be in relation to this issue. The current study creates connective profiles of sermons, statutes and letters over time, and once a historical baseline is established, twenty-first century email and instant messaging. All data are taken from the *Transhistorical Corpus of Written English* (Marcus and Maden-Weinberger 2021).

The connective profile of sermons shows that they consistently demonstrate characteristics of the spoken mode from the fifteenth to the twenty-first centuries. The connective profile of personal letters supports the classification of them as speech-like, although there are more fluctuations in the frequencies of both coordinators and subordinators over time compared to sermons. Statutes, classified as writing-based and purposed, consistently demonstrate characteristics of the written mode over time, such as high frequencies of clause-level *or*. Furthermore, the macro-level results relating to both the personal correspondence and the sermon data support a hypothesis based on previous research which predicts a decrease in oral features from the late Middle English into the Early Modern English period, and a corresponding increase in subordinators marking clauses of cause, condition and concession (CCC-relations), followed by an increase of oral features again moving into the present day, although the hypothesis is somewhat problematized to by micro-level results relating to subordinators marking conditional and causal clauses in sermons. The contrastive diachronic analysis of the historical text types also highlights some potential text-type specific functional motivations for, specifically, the decline of clause-level *and*, especially noticeable in the personal letter data, and the increasing use of *as* as a subordinator marking clauses of reason. The digital text types instant messaging and email were only considered in relation to the twenty-first century sermon

and statute data. Overall, the results relating to them suggest that it is appropriate to class IM as a speech-like text type which exhibits more 'digital orality' (cf Cutler, Ahmar and Bahri 2022) than email.

References

Cutler, Cecelia, May Ahmar and Soubeika Bahri (eds.). 2022. *Digital orality: Vernacular writing in online spaces*. Cham: Springer Nature.

Kohnen, Thomas. 2007. 'Connective profiles' in the history of English texts: Aspects of orality and literacy. In Ursula Lenker and Anneli Meurman-Solin (eds.) *Connectives in the History of* English (Vol. 283). Amsterdam and Philadelphia: John Benjamins Publishing.

Marcus, Imogen and Ursula Maden-Weinberger. 2021. Transhistorical Corpus of Written English. https://www.sketchengine.eu/transhistorical-corpus-of-written-english/.

Comparing null subject use across registers in Singapore English

Gemma McCarley (University of Konstanz)

Despite English's well-known status as a non-null subject language (NNSL), it allows several well-documented exceptions that leave roughly 3% of subjects unexpressed (Torres Cacoullos & Travis 2019): VP-coordination, situational ellipsis, and diary 'pro-drop' (e.g. Haegeman 1990). Setting aside VP-coordination, the vast majority of these null subjects are utterance-initial, leading to the analysis that they are casualties of left-edge deletion (the dropping of weak syllables at the left edge of prosodic phrases) which reconciles these deviations with English's NNSL-hood (Weir 2012). However, global varieties such as Singapore English have attested higher rates of unexpressed subjects, likely an effect of transfer from Chinese (Sato & Kim 2012; Tamaredo 2018). Given the observed subject-to-register variation in American and British English, this seems like fertile ground to investigate if null subject use differs across registers in a variety with unambiguous null subjects.

I used the ICE Singapore corpus (2002) to explore null subject patterns across a wide range of registers and genres (conversations, lessons, broadcasts, speeches, letters, essays, articles, fiction, etc.). A text was chosen from each 'ICE Text Category' (i.e. text type), split into sentences, tokenized, and annotated by hand. Preliminary results are from two texts that are as comparably 'oral' as a spoken and written text can be: a transcription of a spoken conversation (3,111 words; 244 tokens) and a written social letter (876 words; 83 tokens). Orality denotes how representative of speech a text is and has been shown to affect subject expression in Spanish (Walkden et al. 2023). Potentially reflective of this similarity in orality, both texts exhibit a similar rate of null expression (10% and 12%, as shown in Figure 1), nearly four times the rate attested for American English. Although these texts do show that left-edge deletion significantly ($p<0.007$, glm; Bates et al. 2015) accounts for much null realization – most of the unexpressed pronouns occur utterance-initially (Figure 2), there are still plenty of examples in the data that show null subjects following left-edge material, e.g. (1):

(1) "after that Ø took the train then up till Eunos" <ICE-SIN:S1A-001#137:1:A>

Examples like these cannot be accounted for by a left-edge deletion analysis. That being said, even though the overall proportions of null realization are comparable between the two texts, there does seem to be a difference in their left-edge distribution as null subjects occur only utterance-initially in the written text. This distinction ($p<0.005$, Fisher's exact test) may reflect a

split where written Singapore English still adheres to the British English inherited NNSL system in which the 3% of null subjects expressed can be analyzed as left-edge deletion while spoken Singapore English also reflects this possible NSL influence from Chinese. Despite a social letter being as close to casual speech as a written text can reach, the preliminary data suggest that there does appear to remain a distinct enough register difference to merit this difference in behavior. I am in the process of annotating a sample from each text type, accounting for clause-type, coordination, verb-coda, verb-type, verbal inflection, clitics, verb phrase complexity, and negation (cf. Wagner 2016; Schröter 2019). These preliminary results can be confirmed and nuanced once each text type is represented.

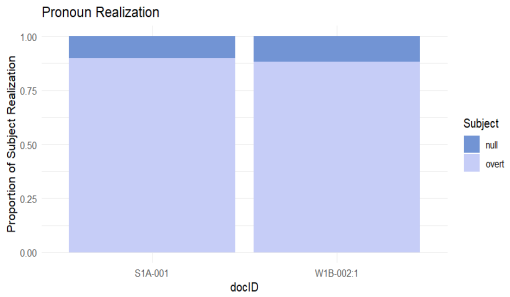


Figure 1. Total proportion of null vs. overt subjects (n=265)

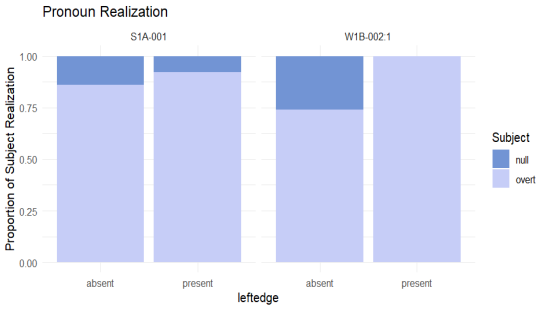


Figure 2. Proportion of null vs. overt subjects against presence of material left of the subject + verb (n=265)

References

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Haegeman, Liliane. 1990. Understood subjects in English diaries: on the relevance of theoretical syntax for the study of register variation. *Multilingua* 9(1), 157–99.

International Corpus of English – the Singaporean Component. 2002. Project Coordinated by Prof. Paroo Nihilani, Dr. Ni Yibin, Dr. Anne Pakir and Dr. Vincent Ooi at The National University of Singapore, Singapore. Available online at <http://ice-corpora.net/ice/download.htm>

The phonology of Nigerian English (PhoNE): A corpus-based acoustic phonetic study of vowels

Philipp Meer (University of Münster)

Work-In-Progress

Nigeria is the most populous country in Africa and boasts an unparalleled linguistic, ethnic, and cultural diversity. Nigerian English (NigE) phonology has been described to have a distinct vowel inventory that incorporates combined vocalic properties of the phonologies of indigenous Nigerian languages and English (e.g. Jowitt 1991, 2019, Gut 2004). Moreover, it has been argued that (i) the Nigerian ethnic groups Hausa, Igbo and Yoruba have different vowel inventories (e.g. Awonusi 1986, Jowitt 1991, 2019), and (ii) systematic differences between speakers from the North and the South exist (e.g. Gut 2004, Brato & Huber 2012, Jowitt 2019). Yet, while research on morphosyntactic, lexical, and pragmatic features of NigE can make use of corpora such as the International Corpus of English (ICE) Nigeria (Wunder et al. 2010), large-scale empirical research on the phonetics and phonology of NigE has remained impossible and is currently restricted to anecdotal observations and small studies (see e.g. Awonusi 1986, Jibril 1986, Jowitt 1991, 2019, Simo Bobda 1995, 2000, 2007, Gut 2004).

By creating semi-automatic phonetic and phonological annotations for ICE Nigeria, the PhoNE project aims to carry out the first large-scale corpus-based investigation of NigE phonology. Within this project, the present paper explores the following research questions:

1) What are the vowels of educated Nigerian English?
2) Are there distinct regional forms of Nigerian English? If yes, what are their properties and how are they geographically distributed?

Using FAVE (Rosenfelder et al. 2014), automatic segmentation and phonemic transcriptions are created and subsequently manually corrected for the following parts of the spoken part of ICE-Nigeria: the broadcast interviews, broadcast news, broadcast discussions and broadcast talks with good audio quality (= ~140,000 words), supplemented by unscripted speeches and non-broadcast talks (= ~50,000 words). Drawing on Bayesian vowel formant estimation for reliable large-scale acoustic analysis (Meer et al. 2021), the study investigates both target-oriented and time-varying acoustic parameters in nominal monophthongs, i.e. vowel inherent spectral change (VISC; Nearey & Assmann 1986). Although VISC has been shown to be important for vowel perception (e.g. Hillenbrand 2013, Morrison & Assmann 2013), phonetic research on postcolonial Englishes has rarely investigated VISC in monophthongs (but see e.g. Meer 2023). Acoustic vowel variation is modeled using established methods such as mixed-effects models, conditional inference trees, random forests, and lectometric techniques (e.g. Tagliamonte & Baayen 2012; Gries 2020; Ghyselen et al. 2020).

While the analysis is currently undergoing, the results will provide large-scale corpus-phonological evidence of the NigE vowel inventory and the extent of systematic regional differentiation in Nigeria.

References

Awonusi, V. 1986. Regional accents and internal variability in Nigerian English: A historical analysis. *English Studies* 6, 555-560.

Brato, T. & Huber, M. 2012. English in Africa. In R. Hickey (ed.), *Areal Features of the Anglophone World*. Berlin: Mouton, pp. 161-185.

Ghyselen, A. S., Speelman, D., & Plevoets, K. 2020. Mapping the structure of language repertoires: On the use of sociolectometric methods. *Zeitschrift Für Dialektologie Und Linguistik* 87(2), 202-249.

Gries, S. T. 2020. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 16(3), 617– 647.

Gut, U. 2004. Nigerian English – phonology. In B. Kortmann & E. W. Schneider (eds.), *A Handbook of Varieties of English: Phonology*. Berlin: De Gruyter, pp. 813-830.

Hillenbrand, J. M. 2013. Static and dynamic approaches to vowel perception. In G. S. Morrison & P. F. Assmann (eds.), *Vowel Inherent Spectral Change*. Berlin: Springer, pp. 9-30.

Jibril, M. 1986. Sociolinguistic variation in Nigerian English. *English World-Wide* 7(1), 47-74.

Jowitt, D. 1991. *Nigerian English Usage: An Introduction*. Ikeja: Longman.

Jowitt, D. 2019. *Nigerian English*. Berlin: De Gruyter.

Meer, P. 2020. Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. *The Journal of the Acoustical Society of America* 147(4), 2283-2294.

Meer, P. 2023. *Standard English in Trinidadian secondary schools: Accent variation and attitudes* (PhD dissertation). University of Münster, Germany.

Meer, P., Brato, T. & Matute Flores, J. A. 2021. Extending automatic vowel formant extraction to New Englishes: A comparison of different methods. *English World-Wide* 42(1), 54-84.

Morrison, G. S. & Assmann, P. F. (eds.) 2013. *Vowel inherent spectral change*. Berlin: Springer.

Nearey, T. M. & Assmann, P. F. 1986. Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America* 80(5), 1297-1308.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H. & Yuan, J. 2014. FAVE (Forced Alignment and Vowel Extraction). Program Suite v1.2.2. https://github.com/JoFrhwld/FAVE.

Simo Bobda, A. 1995. The phonologies of Nigerian English and Cameroon English. In A. Bamgbose, Banjo, A. & Thomas, A. (eds.), *New Englishes: A West African Perspective* Ibadan: Mosuro; Trenton, NJ., pp. 248-268.
Simo Bobda, A. 2000. Comparing some phonological features across African accents of English. *English Studies* 81(3), 249-266.
Simo Bobda, A. 2007. Some segmental rules of Nigerian English phonology. *English World-Wide* 28(3), 279-310.
Tagliamonte, S. A., & Baayen, R. H. 2012. Models, forests, and trees of York English: *Was*/*were* variation as a case study for statistical practice. *Language Variation and Change* 24(2), 135– 178.
Wunder, E.-M., Voormann, H. & Gut, U. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal* 34, 78-88.

Measuring the semantics of English tautological constructions with semantic vector space modeling and sentiment analysis

Qingnan Meng and Martin Hilpert (Dalian Maritime University, University of Neuchâtel)

This research explores the semantics of English tautological constructions (i.e.: "X BE X") used in independent or main clauses, with eight subtypes distinguished by the morphosyntactic features in the X slot. Specifically, we aim to address the following four research questions: 1) What are the most strongly attracted words in slot X of this tautological construction? 2) What is the overall semantic landscape for the "N BE N" construction schema in particular? 3) In general, what sentiment does this construction schema display, and what is the prototypical emotion expressed by each of the 8 sub-construction schemas respectively? 4) What is the semantic difference between "boys" used in tautological construction and its use elsewhere?

In order to answer the three questions above, a corpus-driven quantitative research method is adopted based on the theoretical framework of distributional semantics, with a total of 8,474 concordance lines extracted and then manually checked from Corpus of Contemporary American English (COCA). The whole research procedure is as follows. First, a simple collexeme analysis is conducted to display those words attracted to the X slot with a high collostructional strength. Second, we construct a type-based semantic vector space model to display the overall semantic landscape of this tautological construction schema by visualizing the semantic features of 136 nouns in slot X, and then add contour lines to show the most prominent semantic centers. Third, a lexicon-based sentiment analysis (with "nrc" method) is used to display the sentiment score distribution for all the concordance lines as well as the distribution of eight emotions in Plutchik's (1980) classification over eight sub-constructions. Lastly, a token-based semantic vector space model is constructed for a case study of "boys".

It is found that the top 10 most strongly attracted words are *enough, bygones, boys, rules, kids, facts, the law, business, rape,* and *men* in a descending order, most of which are also the semantic "centers" in the density plot of "N BE N". For the tautological construction with a nominal head in the X slot, its overall semantic landscape is quite miscellaneous, ranging from common noun to proper noun, with various degrees in animacy and abstractness. In general, this tautological construction expresses a slightly positive sentiment, with a prototypical emotion of "trust", followed by "fear", "anger", "sadness" and "anticipation". The semantic differences between the 8 sub-constructions are too subtle to be registered by (at least) American native speakers, which is contrary to the findings in Wierzbicka's (1987) introspective case studies. This is further supported by the case study of "Boys BE boys", in which the meaning of "boys" almost completely overlaps with that of "boys" used elsewhere in the corpus.

References
Fraser, B. 1988. Motor oil is motor oil: An account of English nominal tautologies. *Journal of Pragmatics* 12, 215-220.
Hilpert, M. 2016. Change in modal meanings: Another look at the shifting collocates of *may*. *Constructions and Frames* 8/1, 66-85.
Hilpert, M., & Correia-Saavedra, D. 2020. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory* 16(2), 393-424.
Plutchik, R. 1980. *Emotions: A Psychoevolutionary Synthesis*. New York: Harper & Row.
Wierzbicka, A. 1987. Boys will be boys: 'Radical semantics' vs. 'Radical pragmatics'. *Language* 63, 95-114.
Ward, G., & Hirschberg, J. 1991. A pragmatic analysis of tautological utternances. *Journal of Pragmatics* 15, 507-520.

Syntactic complexity development in intermediate learner English: A longitudinal pilot study

Philine Metzger, Fabian Kettenhofen and Sandra Götz (Philipps-University of Marburg)

Syntactic complexity has featured prominently in Second Language Acquisition research over the last few decades (cf. Larsen-Freeman 2009). Recent developments of tools that can automatically extract a large number of complexity measures (e.g. the *Tool for Automatic Analysis of Lexical Sophistication*; Kyle & Crossley 2015) have led to very detailed descriptions of L2 English complexity development (e.g. Lu 2010; Biber et al. 2011; Kyle & Crossley 2015; Kyle, Crossley & Verspoor 2021). Broadly, we can assume a steadily increasing level of complexity with an increase in learners' proficiency levels, although studies typically report on large degree of variation, so that generalizations are often hard to make. Additionally, despite the comparatively long research tradition in complexity research, truly longitudinal corpus-based studies tracing the complexity development of intermediate learners of English by taking into consideration the effect of learning context variables remain very rare (cf., however, Kyle, Crossley & Verspoor 2021). Studies that not only rely on corpus analyses, but that are complemented by teacher assessments, have – to the best of our knowledge – not been conducted yet.

Against this background, in the proposed paper, we would like to present the findings of a study that investigates how syntactic complexity develops in intermediate German learners of English over four school years while taking into consideration different text types and learning context variables. These findings will be compared to teachers' assessments of the learner texts to check if quantitative complexity measurements correlate with teacher assessments. More specifically, the proposed paper addresses the following research questions:

1) (How) does syntactic complexity develop in written L2 English from grade 9 to grade 12?
2) Do learning context variables have an effect on the development of syntactic complexity of intermediate written L2 English?
3) Are quantitative assessments of syntactic complexity in line with teachers' assessments of learner writing?

In order to answer these research questions, we will analyze a subset of the longitudinal *Marburg Corpus of Intermediate Learner English* (MILE; Kreyer 2015), consisting of written learner data by 90 intermediate learners of English between grade 9 and grade 12, totaling 1,080 texts and more than 500,000 words. In our proposed pilot study, we zoom in closely on 5 learners' developments over 4 years, who submitted 4 texts each year (i.e. 20 essays in total).

These texts were first subjected to an automatic analysis of Lu's (2010) 14 syntactic complexity parameters using the TAASSC tool (e.g. mean length of T-unit, dependent clauses per T-unit, etc.). These variables were also manually analyzed to assess the accuracy of the tool (cf. Châu & Bulté 2022). The data was then subjected to a statistical data analysis using mixed effects regression modelling (e.g. Gries 2015) with the software package R (R core team 2022), while controlling for individual learner variation, differences in text types and learning context variables. One first look into the data suggests that some global complexity variables appear to be robust predictors to discriminate the grade levels, e.g. we observe a steady increase of the mean length of T-units and clauses across grade levels from grade 9-11 (cf. also Larsen-Freeman 1978), whereas we see a decrease from grades 11-12 across some of the investigated variables. Evaluations of the teacher ratings are largely in line with these findings, however, the assessments also revealed some striking differences, which will be discussed in terms of their language-pedagogical implications.

References

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45, 5–35.

Châu, Q. H. & Bulté, B. (2022). A comparison of automated and manual analyses of syntactic complexity in L2 English writing. *International Journal of Corpus Linguistics* 28(2), 232-262, available online https://www.jbe-platform.com/content/journals/10.1075/ijcl.20181.cha

Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10, 95–125.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15, 474–496.

Kreyer, R. (2015). The Marburg Corpus of Intermediate Learner English (MILE). In Marcus Callies & Sandra Götz, eds. *Learner Corpora in Language Testing and Assessment*. Amsterdam: John Benjamins, 13-34.

Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings & application. *TESOL Quarterly* 49(4), 757–786.

Kyle K. & Crossley, Scott A. & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition* 43 (4), 781–812.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly* 12, 439–448.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics* 30(4), 579–589.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Exploring register variation in human and machine-generated texts: A comparative analysis

Jiří Milička[1], Anna Marklová[2] and Václav Cvrček[1]

([1]Charles University, [2]Humboldt University of Berlin)

This study investigates the register variation in texts generated by humans and those produced by transformer based large language models (LLMs, see Vaswani et al., 2017), with a particular focus on the impact of Reinforcement Learning from Human Feedback (RLHF, see Ouyang et al., 2022) on the linguistic diversity in machine-generated texts. LLMs, such as GPT-4, are trained on extensive and diverse datasets and are thus expected to adeptly imitate a wide range of linguistic registers. However, RLHF imposes constraints on the latent space of these models (Casper et al., 2023, see p. 11), often leading to a noticeable limitation in their ability to replicate certain styles and modalities. This limitation aligns with anecdotal evidence, such as attempts to

generate horror narratives using ChatGPT resulting in characters speaking in an unusually polite and formal manner.

Our research investigates whether non-RLHF models, like Davinci-2, outperform current RLHF mod- els (GPT-3.5 Turbo, GPT-4 Turbo), which, despite being larger and trained on more data, might be limited in their capacity to imitate language variability. We employ multidimensional analysis to compare machine-generated texts with those created by humans (Conrad and Biber, 2001; Nini, 2019). Initially, we identify variations in different registers within a corpus using Principal Component Analysis (PCA) of various stylometric, lexical and grammatical features. We then automatically generate continuations of texts from these corpora using various LLMs and subject these to PCA for comparison.

Preliminary analysis, conducted on the Czech and English sections of IntercorpV11§ using Davinci-2, GPT-3.5 Turbo, and GPT-4 Turbo, suggests that both RLHF and non-RLHF models exhibit significantly reduced variability in stylometric variability compared to human-generated texts. Given the rapidly evolving nature of the field, our study will include the latest models available at the time of the conference, such as new LLMs by OpenAI, Mistral, Alphabet, Meta, and other open-source or API-accessible models. This will provide a comprehensive and current understanding of the state of register variation in machine-generated texts. This exploration not only aims to assess the qualities of LLMs but also opens a discussion about the consequences of the fact that LLMs underperform in variability and creativity compared to humans, since it raises concerns that reliance on generative models or even imitating their style by humans could lead to a loss in linguistic variability.

References

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lind- ner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U. & Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. https://arxiv.org/abs/2307.15217

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

Nini, A. (2019). The multi-dimensional analysis tagger. In T. Berber Sardinha, M. Veirano Pinto (eds.) *Multi-dimensional analysis: Research methods and current issues*. Bloomsbury, pp. 67–94.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton F., Miller L., Simens, M. Askell A., Welinder, P., Christiano, P. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35, 27730–27744.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* 30.

Exploring diversity in 'Limitations' discourse: A comparative analysis of English and Spanish research articles in the social sciences

Ana I. Moreno (University of León)

In academic writing, effectively communicating one's own research limitations is a delicate rhetorical task that involves balancing transparency with persuasion. This task is typically described in the discussion or closing section of research articles (RA). Spanish social scientists have reported challenges when articulating limitations in English. This study investigates potential variations in how limitations are presented across English and Spanish-medium journals, aiming to understand the underlying reasons for these challenges. Despite existing research that has studied the generic structure of discussion and/or closing sections in English

(e.g., Cotos et al., 2016), little is known about the local contexts of limitations, their rhetorical effects, and the reasons for the expected variation across English and Spanish.

Adopting intercultural rhetoric (Connor, 2011) and genre-based approaches (Moreno & Swales, 2018), this study identifies the sequences of communicative functions into which limitations are integrated in social science RA discussion and/or closing sections in Spanish and English. Drawing on insights from studies of the "bad news message" in business communication (e.g. Lin, 2020), this study conceptualises limitations as a type of "bad news", surrounded by text segments serving various rhetorical purposes: preparation, mitigation, explication, and reassurance. The study aims to understand how limitations are framed in these sections and explore potential cultural influences.

Two comparable samples of ten RA discussion and/or closing sections each in pedagogy, sociology, psychology, business, and economics were drawn from the Exemplary Empirical Research Articles in English and Spanish (EXEMPRAES) Corpus (Moreno & Swales, 2018). In this corpus, the comparable pairs were matched according to overall topic, study type, audience, and persuasive capacity, and the social science discussion and/or closing sections were annotated for their communicative functions (Moreno, 2021). After identifying the local context of each limitation in the two samples, the present study examines their surrounding segments and reannotates them for their rhetorical purposes. This examination is complemented by interviews with ten authors of the RAs, providing insights into socio-cultural influences.

Spanish social scientists often weave detailed explanations into their limitations, showcasing their expertise and attributing constraints to external factors. On the contrary, English counterparts prefer presenting implications for future work, embedding limitations within positive takeaways. The order of rhetorical purposes also diverges. English authors tend to incorporate more mitigation strategies before stating the limitation, creating nuanced patterns of 'good-bad news'. In contrast, Spanish authors follow more straightforward patterns, emphasising expertise and often placing mitigation strategies after stating the limitation. The study reveals how these variations can be attributed to cultural writing styles, values, and authors' understandings of impression management.

This genre- and corpus-based study contributes to our understanding of the intricate interplay between language, culture, and rhetorical choices in the presentation of research limitations. The findings shed light on why scholars from certain backgrounds may approach this task with distinct concerns. The need for training intercultural competence in higher education, recognising and respecting diverse rhetorical practices, is underscored.

References

Connor, U. (2011). *Intercultural rhetoric in the writing classroom*. University of Michigan Press.

Cotos, E., Link S. & Huffman, S. (2016). Studying disciplinary corpora to teach the craft of discussion. *Writing & Pedagogy* 8(1), 33–64.

Lin, Y. (2020). Communicating bad news in corporate social responsibility reporting: A genre-based analysis of Chinese companies. *Discourse and Communication* 14(1), 22-43.

Moreno, A. I. (2021). Selling research in RA discussion sections through English and Spanish: An intercultural rhetoric approach. *English for Specific Purposes* 63, 1-17.

Moreno, A. I. & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes* 50, 40-63.

Deconstructing economists'arguments: A cross-linguistic study of economic claims through the epistemic and attitudinal stance devices in the bilingual LexEcon corpus

Maria Teresa Musacchio and Dario del Fante (University of Trieste, University of Ferrara)

In the investigation of economic discourse, a major role has been played by the way economists express arguments (Merlini 1983, 1996; Swales 1993; Donohue 2006). The linguistic study of economic argumentation has focused on hedging (Bloor and Bloor 1993) in the discourse of predicting and forecasting (Merlini Barbaresi 1983, 1996, 2005) as specific devices economists use to position their (hybrid) discipline as a soft science increasingly exploiting methods and tools typical of hard sciences. Argumentation has generally been regarded as typical of academic research as expressed in treatises in the past, and in scientific articles more recently. By contrast, textbooks and handbooks are usually regarded as avoiding the presentation of any critical reading since they are supposed to transmit a canon (Swales 1993: 224).

As a research unit within the national project LexEcon – The Economic Teacher: A translational and diachronic study of treatises and textbooks of economics (18th to 20th century) – funded by the Italian Ministry of University and Research, we investigate economic discourse as evidenced by an English, French and Italian corpus of theoretical treatises, textbooks and popular science handbooks published between 1900 and 1970 (45m words) and study economic argumentation in a broad sense, not limited to predicting and forecasting. In this paper, we study economic claims as evidenced by the use of epistemic and attitudinal stance markers – adverbials, complement clauses, nouns + prepositional phrases, and premodifying stance adverbs. These markers are investigated in their role of hedging/boosting features pragmatically expressing degrees of epistemic certainty as opposed to attitudinal inferencing in the construction of economic arguments. Here we focus on an approx. 7m-word subcorpus of English and Italian treatises, textbooks and handbooks published between 1900 and 1929 to take account of the fact that Wall Street Crash is considered a watershed in economics. Triangulating data from previous research (Hyland 2005, 2016; Martin and White 2005; Biber et al. 2021) we have identified lexis that can be regarded as claim-expositive following (and extending) Bloor and Bloor's (1993) typology and studied the epistemic and attitudinal stance markers that go with it. We have then replicated the process to identify claim expositives and concurrent markers in Italian (Gualdo and Telve 2011; Gualdo 2021; 2023) to probe our LexEcon 1900-1929 subcorpus and compare/contrast how claims are presented in treatises to illustrate economic theory and in textbooks/handbooks to train students in argumentation as part and parcel of education in becoming economists.

Preliminary results indicate consistent findings across the two languages, suggesting that claim expositives present co-texts where epistemic and attitudinal stance are primarily conveyed through a core set of adverbs and complement clauses. These patterns exhibit varying frequencies compared to general language corpora such as the BNC and COCA for English, the ItTenTen20 and Paisà for Italian, as well as specialized language corpora, emphasizing the central role of pragmatically hedging/boosting claims. In terms of epistemic certainty in English, the most frequently used adverbs or expressions include 'be possible/possibly/the possibility that', 'be likely', and 'be (un)certain that/to', while 'be obvious/obviously' and 'be/become clear/clearly' are also prevalent, along with 'necessarily'. 'Be sure/surely' ranks highest as an inferencing device. As regards Italian, the most frequent adverbs or expressions are 'è chiaro che, è possibile che/ la possibilità di/la possibilità che/possibilmente, è ovvio che, è probabile che/probabilmente'. The most frequent inferencing devices are 'è certo che/certamente/, è sicuro che/ sicuramente. Future research will extend to cover both the remaining English and Italian component of the 20th-century LexEcon corpus and the 1860-1899 one to provide a more

exhaustive diachronic view of how argumentation as evidenced by lexical expositives and epistemic or attitudinal stance markers evolved in economics up to 1970s.

References

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan E. (2021). *Grammar of Spoken and Written English*. Benjamins.

Bloor, M. and Bloor, T. (1993). How economists modify propositions. In W. Henderson, T. Dudley-Evans and R. Backhouse (eds.), *Economics and Language* (pp. 153-169). Routledge.

Donohue, J.P. (2006). How to support a one-handed economist: The role of modalisation in economic forecasting. *English for Specific Purposes* 25, 200-216.

Gualdo, R. (2021). *Introduzione ai linguaggi specialistici*. Carocci.

Gualdo, R. (2023). *L'italiano dell'economia*. Carocci.

Gualdo, R. and Telve, S. (2011). *Linguaggi specialistici dell'italiano*. Carocci.

Hyland, K. (2005). *Metadiscourse*. Continuum.

Jiang, F.J. and Hyland, K. (2016). Nouns and academic interactions: A neglected feature of metadiscourse. *Applied Linguistics* 4, 1-25.

Martin, J.R. and White, P.R.R. (2005). *The Language of Evaluation. Appraisal in English*. Palgrave MacMillan.

Merlini Barbaresi, L. (1983). *Gli atti del discorso economico: La previsione. Status illocutorio e modelli linguistici nel testo inglese*. Edizioni Zara.

Merlini Barbaresi, L. (1996). Traduzione e pragmatica del discorso. In G. Cortese (ed.), *Tradurre i linguaggi settoriali* (pp. 73-85). Edizioni Libreria Cortina.

Merlini Barbaresi, L. (2005). Il discorso economico/argomentativo: Marcatezza e complessità della previsione. In L. Schena, C. Preite, & S. Vecchiato (eds.), *Gli insegnamenti linguistici nel nuovo ordinamento: Lauree triennali e specialistiche dell'area economico-giuridica* (pp. 301–324). Egea.

Swales, J. (1993). The paradox of value: Six treatments in search of the reader. In W. Henderson, T. Dudley-Evans and R. Backhouse (eds.), *Economics and Language* (pp. 223-239). Routledge.

The concept of 'normal' in the US news discourse during the COVID-19 pandemic

Zuzana Nádraská (Charles University)

This paper contributes to the research on the coronavirus/COVID-19 news discourse (e.g. Cartier et al. 2022; Dong et al. 2021; Fois 2022; Jiang and Hyland 2022; Mattiello 2022; Műller et al. 2021; Nor and Zulcafli 2020; Semino 2021). It explores the meanings and contexts of the adjective 'normal' and shows how 'normality' was conceptualised during the pandemic. The research examines the data from the US section of the Coronavirus Corpus published between January 2020 and December 2022 (Davies 2019-). The News on the Web corpus is used as a reference corpus (Davies 2016-).

The conceptual significance of 'normality' finds reflection in the fact that in the Coronavirus Corpus 'normal' represents the adjective with the highest frequency of occurrence in quotation marks. Moreover, the issue of (the new/old) 'normality' is referred to in various research papers (e.g. Galanopoulos 2020; Jarvis 2021; Sobande and Klein 2022; Zinn 2020). Drawing on the notions of semantic preference and evaluative prosody, the present research focuses on the analysis of clusters (n-grams) and collocation patterns of the examined adjective (Bednarek 2008; Morley and Partington 2009; Partington 2004, 2014; Sinclair 2004; Stewart 2010; Stubbs 2001). Log-likelihood and mutual information measures were applied to assess the significance of the results. The quantitative analysis is supplemented by a qualitative analysis of 200 KWIC lines.

The analysis of collocation serves as the basis for the identification of semantic groups associated with 'normal' (e.g. *return to normal, normal life, temporal specification, degree/range of normality, definition of normal, the existence/appearance of normal*). These categories are also traceable in the identified clusters (e.g. *back to 'normal', a more 'normal', the old 'normal', the new 'normal', semblance of 'normal', 'normal' life, 'normal' times, 'normal' again, 'normal' school year, 'normal' will look*). The data indicate that in the US coronavirus/COVID-19 news discourse in the examined period 'normality' was portrayed as a time-related scalar concept defined by common life activities. The analyses show that 'normal' has developed novel meanings peculiar to the coronavirus/COVID-19 discourse. First, 'normal' refers to the old pre-coronavirus 'normal' and occurs in the context of positive evaluative prosody expressing the hope for the return of the old order. In other contexts, however, (the return to) 'normal' is evaluated negatively as something undesirable deserving re-consideration and change. Second, 'normal' can refer to the post-coronavirus new 'normal' which, though not defined clearly yet, is expected to be qualitatively different from the old 'normal'.

The meanings of 'normal', its semantic preference and evaluative prosody seem to be intricately connected to the presence of quotation marks (cf. Dillon 1988; McDonald 2008; Nádraská 2022; Predelli 2003; Semino and Short 2004). Apart from their emphatic and attention-seeking functions, quotation marks co-signal the unconventionality and contextual dependency of the newly developed meanings, especially in contexts dealing with the conceptualisation of 'normality' (e.g. the semantic groups *degree/range of normality, definition of normal*). Additionally, quotation marks co-indicating the speaker's distance, reservation or disagreement contribute to the expression of negative evaluative prosody. Consequently, the verbal and non-verbal means complement each other to perform identical functions.

References

Bednarek, Monika (2008) Semantic preference and semantic prosody re-examined. *Corpus Linguistics and Linguistic Theory* 4(2): 119-139.

Cartier, Emmanuel, Onysko, Alexander, Winter-Froemel, Esme, Zenner, Eline, Andersen, Gisle, Hilberink-Schulpen, Béryl, Nederstigt, Ulrike, Peterson, Elizabeth and van Meurs, Frank (2022) Linguistic repercussions of COVID-19: A corpus study on four languages. *Open Linguistics* 8: 751-766.

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

Davies, Mark (2019-) *The Coronavirus Corpus*. Available online at https://www.english-corpora.org/corona/.

Dillon, George L. (1988) My words of an other. *College English* 50(1): 68-73.

Dong, Jihua, Buckingham, Louisa and Wu, Hao (2021) A discourse dynamics exploration of attitudinal responses towards COVID-19 in academia and media. *International Journal of Corpus Linguistics* 26(4): 532-556.

Fois, Elenora (2022) News translation and national image in the time of Covid-19. *Brno Studies in English* 48(1): 5-23.

Galanopoulos, Antonis (2020) Return to a new normality? Challenges of the post-pandemic era. *Identities Journal for Politics, Gender and Culture* 17(1): 136-137.

Jarvis, Lee (2022) Constructing the coronavirus crisis: Narratives of time in British political discourse on COVID-19. *British Politics* 17: 24–43.

Jiang, Feng K. and Hyland, Ken (2022) COVID-19 in the news: The first 12 months. *International Journal of Applied Linguistics* 32(2): 241-258.

Mattiello, Elisa (2022) *Linguistic Innovation in the Covid-19 Pandemic*. Newcastle upon Tyne: Cambridge Scholars Publishing.

McDonald, Edward (2008) Maintaining symbolic control over Taiwan: Arguing with scare quotes in the mainland press. In *Communicating Conflict: Multilingual Case Studies of the News Media*, Elizabeth A. Thompson and Peter R. R. White (eds). London: Continuum. 119-141.

Morley, John and Partington, Allan (2009) A few frequently asked questions about semantic – or evaluative – prosody. *International Journal of Corpus Linguistics* 14(2): 139-158.

Müller, Marcus, Bartsch, Sabine, and Zinn, Jens O. (2021) Communicating the unknown: An interdisciplinary annotation study of uncertainty in the coronavirus pandemic. *International Journal of Corpus Linguistics* 26(4): 498-531.
Nádraská, Zuzana (2022) The function of scare quotes in hard news: Metadiscoursal and generic perspectives. *Discourse and Interaction* 15(2): 101-127.
Nor, F. M. and Zulcafli, Adlyn S. Corpus driven analysis of news reports about Covid-19 in a Malaysian online newspaper. *Journal of Language Studies* 20(3): 199-220.
Partington, Alan (2004) Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9: 131-156.
Partington, Alan (2014) Evaluative prosody. In *Corpus Pragmatics: A Handbook*, Karin Aijmer and Christoph Rűhlemann (eds). Cambridge: Cambridge University Press. 279-303.
Predelli, Stefano (2003) Scare quotes and their relation to other semantic issues. *Linguistics and Philosophy* 26: 1-28.
Semino, Elena (2021) "Not soldiers but fire-fighters" - metaphors and Covid-19. *Health communication* 36(1): 50-58.
Semino, Elena and Short, Mick (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London/New York: Routledge.
Sinclair, John (2004) *Trust the Text: Language, Corpus and Discourse*. London/New York: Routledge.
Sobande, Francesca and Klein, Bethany (2023) 'Come and get a taste of normal': Advertising, consumerism and the Coronavirus pandemic. *European Journal of Cultural Studies* 26(4): 493-509.
Stubbs, Michael (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Zinn, Jens O. (2020) 'A monstrous threat': How a state of exception turns into a 'new normal'. *Journal of Risk Research* 23(7-8): 1083-1091.

Explaining American English spellings with reference to the history of British English spellings: Adverbs and prepositions ending in *-ward* and *-wards*

Fujio Nakamura (Kansai Gaidai University)

The form *afterward* is lemmatised prior to *afterwards* in American English (AmE) dictionaries, e.g. *MWCD*[11]. British English (BrE) dictionaries, like *OALD*[10], categorise *afterwards* as typically BrE, while *afterward* is considered the American English variant. Similar descriptions can be found in other contemporary English grammars such as Mencken (1977 [1919]), Quirk et al. (1985) and Huddleston & Pullum (2002). Nevertheless, a notable disparity in usage is evident in corpus analysis. For example, the Corpus of American Soap Operas (2001-2002) contains 136 examples of *afterward*, while *afterwards* dominates with 1,012 examples (respectively, 11.8% and 88.2% of the tokens). However, in other contemporary AmE corpora (listed below), *afterward* yields 13,075 tokens (52.9%), surpassing *afterwards*, which occurs 11,622 times (47.1%). In comparison, in contemporary BrE corpora (see below) *afterward* is attested in 42 examples (0.9%) and *afterwards* in 4,638 examples (99.1%). Since there are few comprehensive studies on the variation between *-ward* and *-wards* forms, this paper aims to explore such variation across time and varieties of English.

Fifteen high frequency words ending in *-ward(s)* have been selected for diachronic and synchronic exploration, including *toward(s)*, *afterward(s), downward(s)* and *forward(s).* The corpora analysed include: (a) contemporary BrE: LOB, FLOB and BNC; (b) contemporary AmE: Time (1923-2006, magazine), Brown, Frown, COCA and SOAP; (c) other varieties of contemporary English: ACE (Australian), Kolhapur (Indian) and Strathy (Canadian); (d) historical BrE: Early English Books Online (EEBO v3) (1470s-1690s), BrE part of ARCHER 3.2 (1600-1999) and Hansard (1803-2005, speeches given in the British Parliament); and € historical AE: AE part

of ARCHER 3.2 (1750-1999), Supreme Court (1790s-2017, US Supreme Court decisions) and COHA.

The evidence gathered so far shows that *afterward* was the preferred form in early BrE. To illustrate this, in EEBO *afterward* far outnumbers *afterwards* in the fifteenth and sixteenth centuries (4,918 tokens for *afterward* (93.4%) vs. 347 for *afterwards* in subperiod 1474-1549 (6.6%), and 18,331 (64.4%) vs. 10,152 tokens (35.6%) in subperiod 1550-1599), though there is clearly a gradual shift towards *afterwards*, particularly evident in the second half of the seventeenth century (24,685 instances of *afterward* (20.0%) vs. 98,440 for *afterwards* (80.0%) in subperiod 1650-1699). The data gleaned from other historical BrE corpora, namely the BrE part of ARCHER and Hansard, point in the same direction. Also, the findings reveal that the older form without *-s* was brought into the US by immigrants and has since been thriving, especially in printed AmE.

References

[*MWCD*[11]] Webster, Noah. [2020[11] (1898)] *Merriam-Webster's Collegiate Dictionary*. Ed. by Marc Bot. Springfield, Mass.: G. & C. Merriam co.

[*OALD*[10]] Hornby, Albert S. 2020[10] [1948]. *Oxford Advanced Learner's Dictionary of Current English*. Ed. by Lea, Diana and Jennifer Bradbery, et al. Oxford: Oxford University Press.

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

Mencken, Henry L. 1977 [1919]. *The American Language: An Inquiry into the Development of English in the United States*. 4th ed., One-Volume Abridged Edition. New York: Alfred A. Knopf.

Quirk, Randolph, Sydney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

A study of the recent evolution of swearing in British teen talk

Ignacio M. Palacios-Martínez (University of Santiago de Compostela)

Swearing has received considerable attention in the literature from a variety of perspectives (cf. Anderson and Trudgill 1990; McEnery 2005; Stapleton 2010; Ljung 2011; Beers and Stapleton 2017; Love 2021; Stapleton et al. 2022, to mention just a few). There are a number of reasons for this: (i) It is an area of interest not only to speakers but also to society in general, in that it is closely associated with the use of bad language and taboo words; (ii) The expression of swearing is not restricted to lexis, and can be found at all linguistic levels; (iii) Following Stapleton et al. (2022), swearing is powerful in the sense that it brings about in the individual a whole set of distinctive psychological, physiological and emotional effects, as well as leading to unique interactional and rhetorical outcomes; and (iv) Teenagers, often seen as key linguistic innovators (Eckert 2014; Tagliamonte 2016), are frequent users of swear and taboo words (Stenström 2006; Stenström et al 2002; Love 2021).

My aim here is to consider the degree to which the swearing practices of teens have changed over time, as has often been observed with other aspects of their language production seen as ephemeral, or whether, on the contrary, differences in the frequency of common swear terms, and more particularly divergences in their use, can be identified diachronically. To this end, I will use a corpus-based approach to explore the behaviour of a set of 12 swear word lemma forms, analysing data drawn from COLT compiled in the 1990s, the LEC (2004-2010), the BNC2014 and Drummonds's material (Drummond 2020).

Preliminary results indicate that differences in terms of the rate of occurrence (normalised frequencies) of the 12 forms under study are recorded across the data of the four corpora. However, these differences in frequency may be attributed to the different methods used in the collection of data i.e., spontaneous productions, group/individual interviews.

Despite divergences in the general frequencies found, no major differences are identified when the total figures are considered in terms of an overall picture. Thus, *fuck*, *shit*, *piss*, *crap* and *bitch* occupy in this order the highest frequencies in all the data sources, while *bastard*, *wank*(*er*) and *cock* all rank, with minor differences, among the least frequent. An exception to this general trend is *bloody*, which seems to be undergoing a clear decline in use, a finding reported elsewhere in the literature on general British English (Love 2021).

All this seems than to indicate that, although the language of teenagers is prone to change quickly, particularly in the area of lexis, this does not appear to be the case with swearing to the same extent.

References

Anderson, Lars and Peter Trudgill. 1990. *Bad language*: Oxford: Blackwell.

Beers, Kristy and Karyn Stapleton (eds.) 2017. *Advances in swearing research: New languages and new contexts*. Benjamins: Amsterdam.

Drummond, Rod. 2020. Teenage swearing in the UK. *English World-Wide* 41(1): 59-88.

Eckert, Penelope. Language and gender in adolescence. In Susan Ehrlich, Miriam Meyerhoff and Janet Holmes, eds. *The handbook of language, gender, and sexuality*. New Jersey: Wiley, 529-545.

Ljung, Magnus. 2011. *Swearing: A cross-cultural linguistic study*. Houndmills, Basingstoke: Palgrave Macmillan.

Love, Robbie. 2021. Swearing in informal spoken English: 1990s–2010s. *Text and Talk* 41 (5-6): 739-762.

McEnery, Tony. 2005. *Swearing in English: Bad language, purity, and power from 1586 to the present*. New York: Routledge.

Stapleton, Karyn. 2010. Swearing. In Miriam A. Locher and Sage A. Graham, eds. *Interpersonal Pragmatics. Handbook of Pragmatics*. Berlin: De Gruyter, 289-305.

Stapleton, Karyn, Knity Fägersten, Richard Stephen and Catherine Loveday. 2022. The power of swearing: What we know and what we don't. *Lingua* 277(7).

Stenström, Anna-Britta. 2006. Taboo words in teenage talk: London and Madrid girls' conversations compared. *Spanish in Context* 3(1): 115-138.

Stenström, Anna-Britta, Gisle Andersen and Ingrid Kristine Hasund. 2002. *Trends in teen talk*. Amsterdam: Benjamins.

Tagliamonte, Sali. 2016. *Teen talk. The language of adolescents*. Cambridge: Cambridge University Press.

Give me a break: The English dative alternation in semantic vector space

Chiara Paolini and Benedikt Szmrecsanyi (KU Leuven)

In this paper, we take a fresh look at an extremely well-studied case of grammatical optionality in language – the dative alternation in English. In English, language users have the choice between two functionally broadly equivalent ways to express dative relations involving a recipient and a theme: the ditransitive dative variant, as in (1), and the prepositional dative variant, as in (2).

(1) But [they]$_{\text{subject}}$ [give]$_{\text{verb}}$ the [guy]$_{\text{recipient}}$ [a job]$_{\text{theme}}$ in prison and make him pay his damn debt. (DAT-2772)

(2) [The judge]$_{\text{subject}}$ [will usually, uh, give]$_{\text{verb}}$ [custody]$_{\text{theme}}$ [to the mother]$_{\text{recipient}}$ ninety-seven percent of the time. (DAT-4067)

The probabilistic conditioning of the dative alternation is in principle well-understood. But much of the literature focusses on traditional, higher-level formal predictors, such as constituent pronominality, constituent definiteness, or constituent length (as in Bresnan et al. 2007). In contrast, semantic properties of the materials in the argument slots have received rather short shrift in the extant literature. The reason is that manually annotating corpus material for top-down semantic predictors (such as constituent animacy, the one semantic predictor typically considered in dative alternation research) is labor-intensive and time-consuming.

Against this backdrop, we present a corpus-based and fully bottom-up method to consider constituent semantics: semantic predictors generated using distributional models of meaning (Lenci 2018). Our research question is the following: How adequately can we predict dative choices as a function of the semantics of dative constituents? So, in example (1), what is the extent to which the theme *job* triggers the ditransitive dative variant? In (2), what is the extent to which the recipient *mother* triggers the prepositional dative variant?

In this spirit, we re-analyze the Switchboard-based US-American section of the publicly available dative alternation dataset available at https://purl.stanford.edu/qj187zs3852 (Bresnan et al. 2017). This dataset is largely identical to the dataset investigated in Bresnan et al.'s seminal (2007) study, and yields N = 1222 observations of the dative alternation after *give*. The semantic vector space models were trained on the spoken COCA (Davies 2008 -, ~127 million words).

Technically speaking, we measure association strengths between the heads of the noun phrases taking the role of theme and recipient and their context words via type-level semantic vector space modeling. Based on the resulting numerical profile, we then cluster both the theme heads and the recipient heads into groupings of semantically-related types and use the resulting clusters as categorical predictors in mixed-effects binary logistic regression analysis.

Results show that recipient heads are clustered into rather coherent groupings related to family roles, job titles, economics and law terminology, as well as anaphoric pronouns. Conversely, theme heads yield a wider range of semantic groupings, including lexemes related to the labor market and household items. Binary logistic regression analysis indicates that while bottom-up semantic predictors have significant predictive power, they are outperformed by traditional predictors, such as constituent weight.

References

Bresnan, J., Cueni, A., Nikitina, T. & Baayen, H. 2007. Predicting the dative alternation, in Bouma, G., Kraemer, I., Zwarts, J. (Eds.) *Cognitive Foundations of Interpretation. Royal Netherlands Academy of Science*, Amsterdam, pp. 69–94.

Bresnan, J., Rosenbach, A., Szmrecsanyi, B., Tagliamonte, S. & Todd, S. 2017. Syntactic alternations data: datives and genitives in four varieties of English. Dataset. *Stanford Digital Repository*. https://purl.stanford.edu/qj187zs3852.

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

Lenci, A. 2018. Distributional models of word meaning. *Annual Review of Linguistics* 4, 151-171.

Understanding fragmental patterns in *let alone* construction: A corpus-based approach

Seulkee Park and Jong-Bok Kim (Kyung Hee University)

The expression *let alone* is typically used after a negative statement to emphasize that the statement also applies even more to the referent of its (bracketed) complement (Harris 2016, Toosarvandani 2008a):

(1) a. Brian would never even read a newspaper, let alone [a book].
b. I hardly have time to think these days, let alone [relax].

In these examples, *let alone* has a remnant complement (*a book* and *relax*) which is associated with its (wavy-underlined) correlate. With these two in a contrastive focus relation, the first clause including the correlate expresses the improbability of a negative statement, and the expression *let alone* plus the remnant at the same time describes a more general, related situation that has not happened, either. Concerning the syntactic and semantic properties of the construction, Harris (2016) and others suggest that the construction is a type of coordination and further derived from move-cum-delete operations, as follows:

(2) Brian would never even read a newspaper, let alone [$_{FocP}$ a book$_i$] <~~Brian would never even read~~ __$_i$>.

The remnant *a book* moves to the focus position, and the remaining clause undergoes ellipsis. This derivation then resorts to the clausal source for the semantic resolution.

We have investigated COCA (Corpus of Contemporary American English) with 1,077 contexts for analyzing quantitative and qualitative data. The dataset explores distributions of remnant fragments and category matchedness with correlates as in Figure 1, which are significant in explaining relations of the pairs.

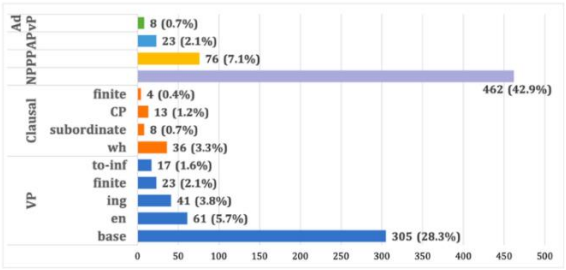


Figure 1: Remnant types and their sub-forms (raw frequencies (%))

Some examples (79 tokens, 7.3%) show category mismatches, which argues against the analysis of the construction as simple coordination:

(3) I haven't had the chance [$_{PP}$ for a break], let alone [$_{VP\text{-}INF}$ to make a phone call]. (2011 SPOK)

In addition, positing clausal sources becomes complicated when the construction appears in the sentence medial position:

(4) A shortage of [fuel] and [lubricating oil], let alone gasoline, would be disastrous to industry. (2008 MAG)

The postulation of clausal sources for such cases requires a cataphoric interpretation, but in real-time processing there is no need to wait until the end of sentences to assign a proper meaning to the construction. A bar graph illustrating the position of remnants with the adjacency to their correlates is presented in the Figure 2.

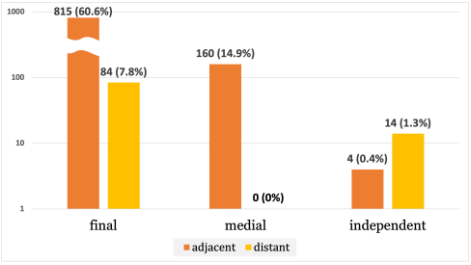


Figure 2.

Attested data like (5) also tell us that we could not assign a negative meaning to *let alone* either:

(5) a. How did you get here, let alone find me?
b. The gaming community needs more people like you, let alone the atheist movement.

The distribution of licensors in Table 1 from our dataset shows that a non-negative environment as in (5) may evoke *let alone* construction very frequently at a rate of 34.3%.

Table 1: The licensing environment of remnants in *let alone* construction

LICENSING ENVIRONMENT	EXAMPLES OF THE LICENSORS	FREQ. (%)
negators	*not, never, no,...*	660 (49.1%)
negative predicates or adverbials	*incapable, unlikely, barely, scarcely, less, yet, bad, difficult, refuse, ...*	224 (16.7%)
quantifiers	*few, most, ...*	16 (1.2%)
non-veridical context	modals, intensional context, ...	368 (27.4%)
non-assertive context	interrogatives, imperatives, conditionals, comparatives, ...	76 (5.7%)
	Total:	1344 (100%)

Unlike Fillmore et al. (1988), we suggest that the construction fits a family of subordination that modifies a non-veridical (non-assertive) situation (Giannakidou 2009). The antecedent clauses in (5) do not have a strong NPI licenser, but depict a non-verdical situation. The coordination-like properties are inherited from the contextually-controlled Parallelism Condition between ellipsis-antecedent (Hartman 2011):

(6) Let-Alone Construction (↑ elliptical-cxt)

The *let-alone* construction, describing a situation *s1*, modifies a nonveridical situation *s0* whose contextual scale is smaller than *s1*.

$$\textit{let-alone-cxt} \Rightarrow \begin{bmatrix} \text{SYN} \mid \text{MOD} \left\langle \begin{bmatrix} \textit{nonverdical} \\ \text{IND } s0 \end{bmatrix} \right\rangle \\ \text{SEM} \mid \text{IND } s1 \\ \text{CNXT} \mid \text{PRESUP} \begin{bmatrix} \textit{cxt-scale } s0{<}s1 \\ \textit{more-prominent}(s0,s1) \end{bmatrix} \end{bmatrix}$$

As implied by this, the *let-alone* XP is interpreted as denoting a situation referring to a discourse. This approach places further contextual constraints with respect to the contextual scale and prominence between the antecedent and the situation evoked from the construction. This

discourse-based direction seems to be more feasible to account for its flexible distributions in real-life situations including dialogues.

References

Fillmore, Charles J., Paul Kay and Mary Catherine O'connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64: 501–538.

Giannakidou, Anatasia. 2009. Negative and positive polarity items: Variation, licensing, and compositionality. In Claudia Maienborn, Klaus von Heusinger and Paul Portner (eds.) *Semantics: An international handbook of natural language meaning*. Berlin: Mouton de Gruyter, pp. 1660–1713.

Harris, Jesse. 2016. Processing *let alone* coordination in silent reading. *Lingua* 169: 70–94.

Hartman, Jeremy. 2011. The semantic uniformity of traces: Evidence from ellipsis parallelism. *Linguistic Inquiry* 42(3): 367–388.

Toosarvandani, Maziar. 2008. Scalar reasoning and the semantics of *let alone*. *Proceedings of Annual Meeting Chicago Linguistics Society* 44(2): 51–64.

Intra-register variation of hedges and boosters in informational-promotional texts: An English-Spanish contrastive study

María Pérez-Blanco and Marlén Izquierdo

(University of León, University of the Basque Country)

Hedges and boosters are recurrent features of the informational-persuasive register of online food descriptions, whose ultimate goal is to both inform and persuade (Biber and Zhang, 2018). As indicators of the writer's epistemic attitude to propositions, they represent two extremes of a cline: hedges express the writer's less than full commitment to the truth of the proposition, whereas boosters emphasise certainty. In relation to the construction of persuasion, boosters help to create an impression of conviction and instill confidence in readers (Vázquez & Giner, 2009). As for hedges, tentative statements would facilitate the audience's acceptability of writer's claims (Hyland, 1996).

Research on hedging and boosting has been largely circumscribed to academic discourse but little extended to other persuasive genres such as newspaper editorials (Dafouz-Milne, 2008; Khabbazi-Oskouei, 2013) or advertising (Fuertes-Olivera et al., 2001; Gustafsson, 2017). Whereas hedging is more researched than boosting in academic discourse (Vázquez & Giner, 2009), the persuasive role of boosters, as markers "accentuating the positive" (Hyland, 2005:78), has been found relevant in advertising (Gustafsson, 2017). The present study aims to identify the most common certainty and uncertainty markers, as they are expected features of the informational-persuasive register. In particular, we are interested in looking into their functions in the different co(n)texts where they appear. Our hypotheses are: i) boosting will outnumber hedging to serve positive evaluation; and ii) the occurrence and functional distribution of each resource will respond to "culturally preferred rhetorical strategies" (Hyland, 2017, p. 25) in promoting a product.

To (dis)confirm our hypotheses, we drew data from ACTEaS_Promo, an English-Spanish comparable corpus of 300 (over 36,000 words) herbal team promotional texts, considered an example of the online food description genre. After examining full texts manually (all 150 in each language), a total of 117 markers in Spanish (ES) and 110 in English (EN) were identified. Among the metadiscourse markers under consideration we found intensifiers ('tons of'), *-ly/-mente* adverbs ('probably', 'notablemente'), periphrases ('puede afectar') or idiomatic expressions ('nothing more, nothing less', 'a manos llenas'). We followed Dafouz-Milne's taxonomy (2008)

(ad-hoc adapted) and paid attention to the functions that hedges and boosters carried out in each functional move.

The findings reveal that, in general terms, boosting exceeds hedging in both languages, especially in EN texts. Most importantly, the data reveal greater diversity in the distribution of the markers in ES, if compared to EN. While hedges and boosters are more evenly spread across moves within the ES texts, the majority of the resources of EN occur in one given move. Looking at this cross-linguistic difference in qualitative terms, we observe that while boosting seems more prominent in moves with a primarily persuasive function, hedging builds moves with a more informational tone. Irrespective of the intra-register variation observed, we conclude that both hedges and boosters play a role in building persuasion in the sort of informational-promotional texts analysed. By skillfully balancing markers that "turn the volume down" (Martin & Rose, 2003) and up, writers confer reliability to the propositional meaning, gaining (the) audience credibility and convincingly presenting the product as attractive and worthy to customers.

References

Biber, D., & Zhang, M. (2018). Expressing evaluation without grammatical stance: Informational persuasion on the web. *Corpora 13*(1), 97-123.

Dafouz-Milne, E., (2008). The pragmatic role of textual and interpersonal metadiscourse markers in the construction and attainment of persuasion: A cross-linguistic study of newspaper. *Journal of Pragmatics* 40(1), 95-113.

Fuertes-Olivera, P., Velasco-Sacristán, M., Arribas-Baño, A. & Samaniego-Fernández, E. (2001). Persuasion and advertising English: Metadiscourse in slogans and headlines. *Journal of Pragmatics* 33, 1291-1307.

Gustafsson, M. (2017). *Metadiscourse in advertising. Persuasion in online advertisements of make up brands*. Unpublished Phd dissertation. Linnaeus University. Sweden.

Hyland, K. (1996). Talking to the academy: Forms of hedging in science research articles. *Written Communication* 3, 251-281.

Hyland, K. (2005). *Metadiscourse. Exploring interaction in writing*. London: Continuum.

Hyland, K. (2017). Metadiscourse: What is it and where is it going*? Journal of Pragmatics* 113, 16-29.

Khabbazi-Oskouei, L. (2013). Propositional and non-propositional: that is the question: A new approach to analyzing 'interpersonal metadiscourse' in editorials. *Journal of Pragmatics* 47, 93-107.

Martin, J.R. & Rose, D. (2003). *Working with discourse: Meaning beyond the clause*. Continuum, London/NY.

Vázquez, I. & Giner, D. (2009). Writing with conviction: The use of boosters in modelling persuasion in academic discourses. *Revista Alicantina de Estudios Ingleses* 22, 219-237.

Out of balance, out of sight: Issues with the design and accessibility of a corpus of fake and real news

Nele Põldvere[1], Zia Uddin[2] and Aleena Thomas[2] ([1]University of Oslo, [2]SINTEF Digital)

Work-In-Progress

Fake news is a topic that only recently has caught the attention of (corpus) linguists (Grieve & Woodfield, 2023; Sousa Silva, 2022; Trnavac & Põldvere, 2024). Such research has sought to identify differences in linguistic features between fake and real news based on carefully designed corpora. An example of such a corpus is the new PolitiFact-Oslo Corpus (Põldvere et al., 2023), a large dataset of fake and real news in English based on recent events (post-2019). However, in its current form the corpus has some limitations, due to the highly specific, and sensitive, nature of fake news. The present methodological study seeks solutions to these limitations with a view to facilitating future corpus building efforts around fake news, a highly promising area of study for linguists.

As the name implies, the PolitiFact-Oslo Corpus relies on the fact-checking website PolitiFact.com for its data, with each news item being individually labelled for veracity by experts (from 'True' to 'Pants on Fire'). In contrast to many other fake news datasets (e.g., DeClarE in Popat et al., 2018), the corpus is the result of a combination of automatic and manual procedures to have greater control over what is included. In addition to a manual approach to text selection, the corpus is accompanied by important metadata information about the texts, such as their text type (e.g., social media) and source (e.g., X). This said, the corpus currently has two major limitations. Firstly, there is a noticeable imbalance between the fake and real news samples (358,516 vs. 70,401 words, respectively), which is due to the preference of PolitiFact and other fact-checkers to debunk false information rather than to find support for true information. This limitation has serious implications for fake news analysis and detection model development based on the corpus (Põldvere et al., 2023). Secondly, due to copyright and privacy issues the corpus is currently not publicly available, a feature of the corpus which is hardly in line with current open science practices.

We offer some solutions. As for the imbalance between the fake and real news samples, we have decided to extend the scope of the fact-checkers rather than to stretch out the timeline. The fact-checkers are found via Google's Fact Check Explorer, which provides quick and easy access to more instances of (mostly or half) true news. The challenge is to ensure comparability of the ratings between the fact-checkers (what is 'Mostly True' according to one fact-checker may be 'Half True' according to another) as well as balance in terms of the metadata information (text type, source). The lack of access to the corpus is a much more complex problem to solve. Inspired by current practices in corpus linguistics, we are exploring opportunities to release the text snippets, rather than the full texts, via an online interface, which, however, is complicated by the legal challenges of distributing fake news data in our national context. We seek solutions to these challenges, too.

References

Grieve, J. & Woodfield, H. (2023). *The language of fake news.* Elements in Forensic Linguistics. Cambridge University Press. https://doi.org/10.1017/9781009349161

Popat, K., Mukherjee, S., Yates, A. & Weikum, G. (2018). DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 22–32). Association for Computational Linguistics.

Põldvere, N., Uddin, Z. & Thomas, A. (2023). The PolitiFact-Oslo Corpus: A new dataset for fake news analysis and detection. *Information* 14, 627. https:// doi.org/10.3390/info14120627

Sousa-Silva, R. (2022). Fighting the fake: A forensic linguistic analysis to fake news detection. *International Journal for the Semiotics of Law* 35, 2409–2433. https://doi.org/10.1007/s11196-022-09901-w

Trnavac, R. & Põldvere, N. (2024). Investigating Appraisal and the language of evaluation in fake news corpora. *Corpus Pragmatics*. https://doi.org/10.1007/s41701-023-00162-x

---

Expressing prediction and forecast across languages: Some insights from the LexEcon corpus

Carla Quinci (University of Padova)

Economics has been generally considered a soft (vs hard) discipline due to its relative inability to predict (cf. Shapin 2022). Yet, formulating hypotheses and scenarios and making predictions constitute the distinguishing feature of the economic discourse (Merlini Barbaresi 2005, 309; Musacchio 1995, 17–18), especially since the early 1900s, when statistics and mathematics started to be increasingly implemented in economic forecasting. In her pivotal analysis of the verbs expressing prediction and forecasting in Economics, Merlini Barbaresi (1983, 1984, 2005)

observed the different nature of economic predictions, which can be interpretative, illustrative, applied, and instrumental, and can be differently positioned on the epistemic gradient and the inferential continuum (Merlini Barbaresi 1983) depending on, respectively, the levels of certainty and subjectivity they convey. Semantics, tense, and modality were found to be one of the main devices carrying epistemic and inferential value in the economic discourse (Musacchio 2017, 57; Donohue 2006; Merlini Barbaresi 1984).

In the attempt to complement and expand existing research in this field, this study offers a contrastive interlinguistic analysis of the verbal expressions used for formulating predictions and forecasts in 46 English and 26 Italian Economics textbooks and handbooks published between 1900 and 1929 (12.5 million tokens). These represent two subcorpora of the much larger *LexEcon* corpus, a diachronic and multilingual collection of 18th-, 19th- and 20th-century volumes in six different languages (Italian, French, English, German, Spanish and Portuguese). The analysis focuses on the frequency and use of 32 lexical and (semi-)modal English verbs (Biber et al. 2021, 482–96) and 36 lexical and modal Italian verbs selected by drawing on previous research and expanded by including the synonyms and the verbal instances concerning prediction/forecasting and modality found among the first 200 verbs of the respective subcorpora. These cover a large spectrum of the epistemic gradient and the inferential continuum as they include assertive (e.g. *believe*, *think, find, pensare, ritenere, credere*), predictive (e.g. *assume*, *predict, expect, prevedere, predire, attendere*), and modal verbs, as well as the semi-modals *be going to* and *need to*. The analysis suggests a preponderance of specific predictive verbs (i.e. *suppose*, *assume, prevedere*), which are purposely used when opposing scientifically grounded predictions to guesswork (e.g. *guess, bet, scommettere, profetizzare*). The major role of prediction in the economic discourse is testified by the high incidence of future and conditional modalities in both languages, which are employed in a wide range of predictions, with generally high epistemic gradients. The modality of possibility, having a lower epistemic gradient, is largely present in both subcorpora and appears to be used in 'irrefutable predictions', i.e. when two opposite outcomes are simultaneously envisaged. Despite some exceptions mostly due to systemic asymmetries between English and Italian, the two corpora seem to be largely aligned with reference to predictive verbal instances, which suggests a common conception of and approach to economic forecasting irrespective of the language involved.

References

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

Donohue, James P. 2006. How to Support a One-Handed Economist: The Role of Modalisation in Economic Forecasting. *English for Specific Purposes* 25 (2): 200–216. https://doi.org/10.1016/j.esp.2005.02.009.

Merlini Barbaresi, Lavinia. 1983. *Gli atti del discorso economico: status Illocutorio e modelli nel testo inglese: la previsione*. Parma: Edizioni Zara.

Merlini Barbaresi, Lavinia. 1984. Will, may e l'espressione della supposizione e dell'inferenza. In *Letteratura e deduzione & Discourse Analysis. Atti del VI Convegno Nazionale dell'Associazione Italiana di Anglistica, Pavia 22-23-24 Ottobre 1983*, edited by Tomaso Kemeny, Lia Guerra, and Anthony Baldwin, 123–29. Fasano: Schena.

Merlini Barbaresi, Lavinia. 2005. Il discorso economico/argomentativo: marcatezza e complessità della previsione. In *Gli insegnamenti linguistici nel nuovo ordinamento: lauree triennali e specialistiche dell'area economico-giuridica*, edited by Leandro Schena, Chiara Preite, and Sara Vecchiato, 301–24. Milano: Egea.

Musacchio, Maria Teresa. 1995. *La traduzione della lingua dell'economia dall'inglese all'italiano*. Trieste: Edizioni LINT.

Musacchio, Maria Teresa. 2017. *Mediating across Languages and Cultures: Economics and Finance as Popular Science in Translation*. Padova: CLEUP.

Shapin, Steven. 2022. Hard Science, Soft Science: A Political History of a Disciplinary Array. *History of Science* 60 (3): 287–328. https://doi.org/10.1177/00732753221094739.

If-clauses in Early and Late Modern English

Nicolás Raths (Johannes Gutenberg University of Maniz)

If-clauses express states of affairs in which a proposition in the main clause is contingent on the fulfilment of a condition in the subordinate clause (cf. Quirk et al: 1088). The level of likelihood that the condition is met is encoded in the morphosyntax of such conditional constructions. And these morphosyntactic configurations gave rise to three or four traditionally distinguished *if*- clause types, as we find them, for instance, in school grammar books (cf. Maloney et al. 2018: 155). The present paper sets out to investigate the distribution of attested *if*-clause types in diachronic perspective. First results indicate that we find several *if*-clause types that violate the morpho-syntactic configurations provided in traditional typologies (cf. also Gabrielatos 2013: 156). Moreover, we observe that diachronically, the *if*-clause types differ considerably regarding their frequencies between 1500 and 1900.

The empirical analysis traces the diachronic development of *if*-clauses in three historical corpora, i.e. *Early English Prose Fiction*, *Eighteenth-Century Fiction* and *Nineteenth-Century Fiction*. The corpora contain British prose fiction published during the Early Modern English (1500-1700) and Late Modern English (1700-1900) periods. This selection of corpora permits us to draw on a sizeable amount of data while keeping the genre constant. The study aims at answering four research questions:

1) Does the overall frequency of conditional *if*-clauses change over time?
2) Does the relative frequency of the four *if*-clause types change over time?
3) Does the frequency of subjunctive verb forms in *if*-clauses change over time?
4) Does the use of modal auxiliaries in *if*-clauses change over time?

1000 occurrences of the subordinator *if* were sampled for each of the three time-spans investigated: 1500-1700, 1700-1800 and 1800-1900. Each occurrence was manually edited in order to exclude false positives (such as concessive *as if*, interrogative *ask if* and *if* in comparative complements *than if*). Preliminary results reveal that the relative frequencies of conditional *if* vs. non-conditional *if* do not fluctuate significantly over time, i.e. the proportion to which the use of *if* functions as a conditional clause has remained relatively constant. However, if we look at the different *if*-clause types, we observe striking fluctuations and considerably more *if*-clause types than traditional four-fold typologies suggest. Annotating the hits according to type, based on the verb phrase in the subordinate and the main clauses, reveals that the frequency of type 3 *if*-clauses appears relatively stable over time, ranging between 5-7%. Interestingly, Type 2 is most frequent in the 17th and 18th centuries (23% and 30%) but then declines in use in the 19th century (12%). The present paper will suggest that this finding can at least partly be attributed to the significant rise in the use of modal verbs during the Modern English period. Type 1 is most frequent in the 18th century, accounting for 1/3 of all *if*-conditionals. Finally, the use of Type 0 remains relatively low in the first two centuries (4% and 5%), but gains ground in the 19th century (15%).

References

Early English prose fiction (1997–2015). Ed. by Holger Klein, David Margolies & Janet Todd. Chadwyck-Healey. ProQuest LLC.

Eighteenth-century fiction (1996–2015). Ed. by Judith Hawley, Tom Keymer & John Mullan. Chadwyck-Healey. ProQuest LLC.

Gabrielatos, Costas (2013). *If*-conditionals in ICLE and the BNC: A success story for teaching or learning? In Granger, Sylviane; Gilquin, Gaëtanelle & Fanny Meunier (eds.) *Twenty Years of Learner Corpus*

Research: Looking back, Moving ahead (Corpora and Language in Use – Proceedings 1), 155-166. Louvain-la-Neuve: Presses universitaires de Louvain.
Gabrielatos, Costas (2021). If-conditionals: Corpus-based Classification and Frequency Distribution. *ICAME Journal* 45(1): 87–124.
Maloney, Paul; Ringel-Eichinger, Angela & Geoff Sammon (2018 *Englische Grammatik für die Mittel- und Oberstufe*. Berlin: Cornelsen.
Nineteenth-century fiction (1999–2000). Ed. by Danny Karlin & Tom Keymer. Chadwyck- Healey. ProQuest LLC.
Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey & Jan Svartvik (1985) *A Comprehensive Grammar of the English Language*. London/New York: Longman.
Traugott, Elizabeth Closs, Alice ter Meulen, Judy Snitzer Reilly & Charles A. Ferguson (1986). *On Conditionals*. Cambridge: Cambridge University Press.

'I make them drink heartily of warm Water three or four Times': Periphrastic causative constructions in Late Modern English scientific writing

Jesús Romero-Barranco (University of Málaga)

Periphrastic causative constructions or analytic causative constructions could be defined as "two-part configurations such as *He makes me laugh* or *I had my hair cut*, where a causative verb controls a non-finite complement clause and which express a causal relation in which the occurrence of the effect is entailed" (Gilquin 2010: 1; Wolff and Song 2003). According to Gilquin, different structural patterns have different distributions (from the very frequent [X *make* Y V_{INF}], [X *get* Y V_{PP}] and [X *have* Y V_{PP}]; to the extremely rare [X *make* Y V_{PP}], [X *have* Y V_{PRP}] and [X *have* Y V_{INF}]), a fact which may depend on the characteristic features of specific text types, the structure [X *cause* Y $V_{TO\text{-}INF}$] being "typical of scientific and technical genres [due to a] higher proportion of nominal (rather than pronominal) elements" (2010, 277). These structures could be understood as a chain in which the energy is transmitted from one entity to the next, as in *Fear*$_{CAUSER}$ *caused John*$_{CAUSEE}$ *to kill*$_{EFFECT}$ *the burglar*$_{PATIENT}$ (Langacker 2002: 254).

In the literature, Stocker (1990) and Hollmann (2000, 2003) assessed the phenomenon from a diachronic perspective, and Talmy (1986), Kemmer and Verhagen (1994), Stefanowitsch (2001) and Gilquin (2010) did so following a cognitive approach. Moreover, Cottier (1991) focused on *cause*, *get*, *have* and *make*; Ikegami (1989, 1990a, 1990b) studied the use of *have* and *get*; and Kemmer (2001) analysed *make*. Apart from these approaches, as far as I have been able to investigate, the quantitative (distribution of different causative verbs and their competition over time) and qualitative (different meanings and structural patterns) aspects of the phenomenon have not been studied in Late Modern English scientific writing so far. The present study pursues, therefore, the following objectives: 1) to study the distribution of *cause*, *get*, *have* and *make* in causative constructions in Late Modern English scientific writing; 2) to assess the different levels of attestation in the different text types in the corpus (from medical recipe collections to scientific periodicals, among others; 3) to analyse the different structural patterns in causative constructions over time; and 4) to provide the typology of verbs that have been found to occur in the causative constructions, i.e. 'I make them drink heartily of warm Water three or four Times'. The source of evidence comes from *The Corpus of Late Modern English Medical Texts* (Taavitsainen and Hiltunen 2019), a 2-million-word corpus covering the period 1700-1800, whose textual division will allow for the detection of different tendencies in different text types over time.

References
Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Grammar of Spoken and Written English*. London: Longman.
Cottier, E. 1991. Les operateurs causatifs de l'anglais: MAKE, CAUSE, HAVE et GET. *Cahiers de recherche en grammaire anglaise* 5: 85–126.
Gilquin, Gaëtanelle. 2010. *Corpus, Cognition and Causative Constructions*. Amsterdam: John Benjamins.
Hollmann, Willem B. 2003. *Synchrony and Diachrony of English Periphrastic Causatives: A Cognitive Perspective*. PhD dissertation, University of Manchester.
Ikegami, Y. 1989. 'HAVE + object + past participle' and 'GET + object + past participle' in the SEU Corpus. In *Meaning and Beyond. Ernst Leisi zum 70*, edited by Geburtstag, U. Fries and M. Heusser. Tübingen: Gunter Narr. 197–213.
Ikegami, Y. 1990a. 'HAVE/GET/MAKE/LET + object + (to-)infinitive' in the SEU Corpus. In *Bumpo to Imi no Aida: Kunihiro Tetsuya Kyoju Kanreki Taikan Kinen Ronbunshu (Between Grammar and Meaning: A Festschrift for Professor Tsuya Kunihiro)*, edited by S. Tsuchida et al. Tokyo: Kuroshio Shuppan. 181–203.
Ikegami, Y. 1990b. 'HAVE + object + V-ING' and 'GET + object + V-ING' in the SEU Corpus. In *Annales Universitatis Scientiarum Budapestinensis De Rolando E.tv.s Nominatae. Sectio Linguistica. Tomus XXI*, edited by I. Szathmári. 93–108.
Kemmer, S. 2001. Causative constructions and cognitive models: The English make causative. In *The First Seoul International Conference on Discourse and Cognitive Linguistics: Perspectives for the 21st Century*. Seoul: Discourse and Cognitive Linguistics Society of Korea. 803–846.
Kemmer, S. & Verhagen, A. 1994. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5(2): 115–156.
Langacker, R. W. 2002. *Concept, Image, and Symbol. The Cognitive Basis of Grammar*. Berlin: Mouton de Gruyter.
Stefanowitsch, Anatol. 2001. *Constructing Causation: A Construction Grammar Approach to Analytic Causatives*. PhD dissertation, Rice University.
Stocker, M. 1990. *The Causative in Middle and Modern English*. PhD dissertation, Universität Basel.
Taavitsainen, I. and T. Hiltunen. 2019. *Late Modern English Medical Texts*. Amsterdam: John Benjamins Publishing Company.
Talmy, L. 1986. Force dynamics as a generalization over 'causative'. In *Georgetown University Round Table on Languages and Linguistics 1985*, edited by D. Tannen and J. E. Alatis. Washington DC: Georgetown University Press. 67–85.
Wolff, P. & Song, G. 2003. Models of causation and the semantics of causal verbs. *Cognitive Psychology* 47: 276–332.

'I hate complaining': A corpus-based study of complaint metadiscourse in the Early and Late Modern English period

Sofia Rüdiger (University of Bayreuth)

Complaints are ever-present in our speech activities, be it when making small talk about the weather (which we find too cold), chatting about politics (which we disagree with), or engaging with the waiter in a restaurant about our food (which was too salty). Third-party complaints in particular (i.e., complaining to someone who is not held responsible for the complainable) are ubiquitous in everyday communication (Boxer 1993: 110). However, despite their many important functions in building rapport (Boxer 1993) and eliciting 'emotional reciprocity' (Günthner 1997), third-party complaints are often socially stigmatized (e.g., Heinemann & Traverso 2009: 2381). This talk sets out to add a historical corpus linguistic perspective on

complaint metadiscourse, i.e., displays of reflective awareness (Haugh 2018) as performed in people's speech or writing (cf. Jucker 2020).

To do so, this study draws on the *Corpus of Early English Correspondence* (CEEC) and its extension (CEECE) (see, e.g., Raumolin-Brunberg & Nevalainen 2007, Laitinen 2002 and https://varieng.helsinki.fi/CoRD/corpora/CEEC/index.html). Taken together, the corpora of 15-18th century British English letters consist of ca. 4.7 million words. The corpora were searched for common complaint terminology (compiled with *The Historical Thesaurus of English* and including spelling variants). The hits thus generated (>1,000) were then manually disambiguated and coded regarding the type of complaint that they concern (i.e., direct vs. third-party vs. ambivalent). In all cases, complaint terminology can be used either 1) to introduce a complaint (i.e., as illocutionary force indicating device), 2) to give descriptive statements about an individual's speech behavior, or 3) to make a metadiscursive comment (e.g., "I hate complaining. Tis no sign I am easy, that I do not trouble you with my Headachs and my spleen. To be reasonable one should never complain but when one hopes redresse." CEEC_MONTAGU_045_1712).

Focusing on third-party complaints specifically, collocate analysis revealed that letter writers displayed a predominantly negative stance towards the act of complaining. As further qualitative analysis shows, the legitimacy of the complaints as well as their quantity was particularly relevant for these judgments. The data also revealed some indications of positive functions for the complainer (i.e., relief of the complainable) and for social relations (i.e., maintenance of friendship). This can also be contrasted via keyword analysis to the direct complaints, which are presented as much more formal and institutionalized (as reflected in the use of lexical items such as 'lord' or 'king'). As an outlook and to corroborate the findings, I will present first results of a similar study using the UK Etiquette Books Corpus which subsumes 28 texts from the 19th century (Paternoster 2022). Further research could also include other text types, such as letter-writing manuals and fictional texts.

References

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

Günthner, Susanne. 1997. Complaint Stories: Constructing Emotional Reciprocity among Women. In Helga Kotthoff & Ruth Wodak, eds. *Communicating Gender in Context*. Amsterdam/Philadelphia: John Benjamins. 179-218.

Haugh, Michael. 2018. Corpus-based Metapragmatics. In Andreas H. Jucker, Klaus P. Schneider & Wolfram Bublitz, eds. *Methods in Pragmatics*. Berlin/New York: De Gruyter. 619-644.

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

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

Laitinen, Mikko. 2002. Extending the Corpus of Early English Correspondence to the 18th Century. *Helsinki English Studies* 2: n.p.

Paternoster, Annick. 2022. *Historical Etiquette – Etiquette Books in Nineteenth Century Western Cultures*. Cham: Palgrave Macmillan.

Raumolin-Brunberg, Helena & Terttu Nevalainen. 2007. Historical Sociolinguistics: The Corpus of Early English Correspondence. In Joan C. Beal, Karen P. Corrigan & Hermann L. Moisl, eds. *Creating and Digitizing Language Corpora. Volume 2: Diachronic Databases*. Cham: Palgrave Macmillan. 148-171.

Register-specificity, polysemy and the atomic metaphor of extended units of meaning

Mathias Russnes (University of Oslo)

This paper investigates the extent to which semantic prosodies are specific to particular registers. The empirical focus of the study is on extended units of meaning which have the polysemous cores *possession, strike, penalty* and *the edge of*, chosen for their distinctive meanings inside and outside of a football context. Semantic prosody describes meaning that "can be seen to reside not simply in the word" itself, but rather belonging to a larger unit that consists of a core and its collocates (Partington 1998: 67). This model of extended unit of meaning can be exemplified through polysemous items, in that their meaning is ambiguous in isolation, and therefore context-dependent (Rundell 2018: 7). The prosodies of certain extended units have also been shown to vary across registers (Hoey 2004; Xiao & McEnery 2006; Hunston 2007; Ebeling 2021), although this has generally not been viewed in relation to polysemy in previous research. In this study, material from the English part of the English-Norwegian Match Report Corpus (ENMaRC) will be compared with data from the British National Corpus 2014 (BNC2014), representing registers, fiction and newspaper texts. The study addresses the following research question: to what extent are the prosodies of extended units with polysemous cores register-specific?

The preliminary results of the study suggest that the prosodies of units can vary across registers, and that this can be connected to the separate senses of the units' cores. To visualise this register-specificity, a novel approach that draws on the structure of atoms as a metaphor will be taken. This approach can be illustrated by the following example, which represents the prosodic environment of *possession* in newspaper texts in the BNC2014:

The item's most prevalent collocates, shown in the figure, indicate the most prominent senses occurring in these registers, as well as their accompanying evaluations, which can be viewed as symptoms of their prosodies. The collocates in green are connected to the lexical item in a sport sense, i.e. having control of the ball, and derive from registers discussing this semantic field, where *possession* has a predominantly favourable prosody, correlating with its use in the ENMaRC, e.g (1). This prosody is contrasted by that of its use in a literal sense, which is clear and negative, e.g (2), and has a strong collocation with lexical items indicating negative evaluative meaning. These are shown in orange in the figure, and derive from news reportage registers. However, when comparing these results with the item's occurrences in fictional texts in the BNC2014, the picture is altered. Here, the sport sense has disappeared, and the literal sense appears to lose its negative evaluation, taking on a more unclear prosody. In this register, the lexical item is also frequently used in a metaphorical sense with immaterial objects, a sense that often expresses evaluation, most commonly with negative meanings, e.g. (3).

(1) Burnley enjoyed their share of *possession* and forced City into making mistakes (ENMaRC)
(2) He was charged with *possession* of heroin (BNC2014)
(3) they appeared not to belong to me, but to a stranger who had taken *possession* of my body (BNC2014)

References

Ebeling, S. O. 2021. Hope for the future: An analysis of HOPE/HÅP(E) across genres and languages. *Bergen Language and Linguistic Studies* 11(1), 7-26.

Partington, A. 1998. *Patterns and Meanings. Using Corpora for English Language Research and Teaching*. John Benjamins Publishing Company.

Hoey, M. 2004. The textual priming of lexis. In G. Aston, S. Bernardini & D. Stewart (eds.) *Corpora and Language Learners*, pp. 21-41. John Benjamins Publishing Company.

Hunston, S. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2), 249-268. John Benjamins Publishing Company.

Rundell, M. 2018. Searching for extended units of meaning – and what to do when you find them. *Lexicography ASIALEX* 5, 5-21. Springer.

Xiao, R. & McEnery, T. 2006. Collocation, semantic prosody, and near synonymy. A cross-linguistic perspective. *Applied Linguistics* 27(1): 103-129. Oxford University Press.

A corpus-based approach to human and machine translation using multidimensional analysis

María del Mar Sánchez-Ramos and Muhammad Shakir

(University of Alcalá, University of Münster)

Work-In-Progress

Application of computational techniques and advanced statistical methods, such as multidimensional analysis (MDA) as developed by Douglas Biber (1988), is one of the most innovative and fruitful lines of research into the characterization of registers based on the co-occurrence of lexico-grammatical patterns (linguistic features). Although MDA has become consolidated in linguistic studies (Biber 1991; Biber and Finegan 1994a, 1994b, Parodi, 2007), its incorporation in translation studies is at an early stage (Calzada Pérez and Sánchez Ramos, 2022; Chou and Liu, 2023; Kruger and Van Rooy, 2016). These studies are even less numerous if human translation (HT) and machine translation (MT) are compared.

The main objective of our research is to incorporate a translational perspective into Biber's work and to shed some light on the main similarities and differences between human translated and machine translated texts. Our data includes 158 text files (832,950 tokens) drawn from the parallel corpus (English-Spanish) EUCJ comprising judgments referring to Spanish Courts and delivered by the European Union Court of Justice (Vigier Moreno and Sánchez Ramos, 2017). These files are translated by humans (HT) and by the neural machine translation system MTUOC-translator (Oliver, 2021). We use the MFTE tagger (Le Foll and Shakir 2023) to obtain per-hundred-word frequencies of more than 100 lexico-grammatical features based on the work of Biber (1988, 2006) and Biber et al. (1999). We then apply principal component analysis (PCA) in the R programming language to get seven components or dimensions of variation. The final solution consists of 76 lexico-grammatical features with a KMO of 0.73 (middling). The 7-component solution explains 43% variance.

Our exploratory research highlights that Dim2, Dim3, and Dim5 are the dimensions that show significant differences in the mean dimension scores of HT versus MT, where HT always has a higher score in all three dimensions. Dim2 mainly consists of verbal (e.g. activity verbs, nonfinite verb *-ing* forms, adverbs), clausal (e.g. *WH* complement clauses, non-finite present participial relative clauses), and informal features (e.g. discourse/pragmatic markers, verb particles) versus nominal (e.g. determiners, prepositions, cognitive nouns, technical nouns, human nouns) and narrative features (e.g. third person pronouns, communication verbs, past tense, perfective aspect) on the negative pole. Dim3 includes many features included on the positive side of Dim2 along with the past tense and perfective aspect (positive pole) as compared to noun and adjective dependent features (attributive adjectives, *that* relative clauses, *to* clauses after evaluative adjectives etc.). Dim5 also includes verbs (e.g. facilitation and causation verbs) and stance-related features (prepositions preceded by stance nouns, *to* clauses preceded by stance nouns) on the positive side versus adverbs (place) and adjectives (relational) on the negative side.

While by no means conclusive, these preliminary results show that HT presents more complexity in terms of syntactic and verb tense structure than MT, which is characterized by a more homogeneous syntactic construction. These findings are in line with previous results described in Vassenhove et al. (2019, 2023), which also indicate a loss of lexical richness in machine translated texts when compared to human-generated texts. In this line, these findings offer a starting point for future research on the existing debate on the categorization of machine translated language as "genre" of its own.

References

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Biber, D. 1991. Oral and literate characteristics of selected primary school reading materials. *Text* 11: 73-96.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education.

Biber, D. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. John Benjamins.

Biber, D. and Finegan, E. 1994a. Multi-dimensional analyses of authors' style: Some case studies from the eighteen century. In D. Ross and D. Brink (eds.), *Research in Humanities Computing 3*, (pp 3-17). Oxford University Press.

Biber, D. and Finegan, E. (eds.) 1994b. *Sociolinguistic Perspectives on Register*. Oxford University Press.

Calzada Pérez, M. and Sánchez Ramos, M.M. 2022. MDA Analysis of translated and non-translated parliamentary discourse. In M. Ji and M. Oakes (eds.), *Corpus Exploration of Lexis and Discourse in Translation* (pp. 26-55). Routledge.

Chou, I. & Liu, K. 2023. Style in speech and narration of two English translations of Hongloumen: A corpus-based multidimensional study. *Target* 36(1), 76-111.

Kruger, H. and Van Rooy, B. 2016. Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide: A Journal of Varieties of English* 37(1): 26–57. https://doi.org/10.1075/eww.37.1.02kru

Le Foll, E. and Shakir, M. (2023). MFTE Python (Version 1.0) [Computer Software]. https://github.com/mshakirDr/MFTE

Oliver, A. 2021. MTUCO-translator. [Computer Software]. https://github.com/aoliverg/MTUOC-translator

Parodi, G. 2007. Variation across registers in Spanish: Exploring El Grial PUCV Corpus. In G. Parodi (ed.), *Working with Spanish Corpora* (pp.11–53). Continuum.

The use of stance markers across sections of BA theses by L1 Estonian learners of English: The effect of sections as sub-registers

Denys Savchenko (University of Tartu)

Work-In-Progress

I will examine the use of stance devices in the written language of L1 Estonian learners of English in comparison with L1 English expert academic writers. Following Biber et al. (1999) and Biber (2006), stance markers are defined as linguistic means of expressing attitudes and/or commitment towards propositions. Larsson (2019) showed that, compared to expert writers, learners rely more on stance adverbs and attitudinal markers of stance. Thus, I will take into account syntactic (stance adverbs: e.g *possibly*, nouns controlling a *that*-clause or followed by a prepositional phrase: e.g. *fact that*, *possibility of*, complement clause constructions: e.g. *it is possible that*) and semantic categories (certainty, likelihood and attitude) of stance. I predict that learners will differ from expert writers in terms of both of these categories.

In view of register as a functional category related to language use in a specific context (Halliday, 1978), previous studies (Biber 2006, Biber and Conrad 2019) showed the interconnection between syntactic realisations of stance and register. Moreover, Larsson (2019) demonstrated that register is an important factor to consider in learner language studies as learners exhibit a certain degree of register unawareness.

In light of previous studies, I will analyse the use of stance markers in a corpus of BA theses, collected at Tartu University, in comparison with the BNC (British National Corpus, Burnard 2007). Larsson (2019) focused on the differences in syntactic realisations of stance in registers in BNC and learners' theses. However, the question of difference in the use of stance markers according to theses sections has not been addressed. As register in a given text can be defined at different levels of specificity (Biber and Conrad, 2019), several recent studies have turned attention to the importance of within-register variation (Egbert & Mahlberg, 2020, Egbert & Gracheva, 2022). Thus, I will focus on three main sections in BA theses and academic articles: introduction, literature review and empirical analysis. Additionally, I will take into account the discipline of the theses (linguistics, literature), status (novice and expert writers), semantic and syntactic categories.

The goal of the study is to examine which factors will have an effect on the distribution of stance markers and whether sub-register should be considered to increase comparability in future studies. For this purpose, I will use Mixed-effect regression modelling (see Gries 2021) to examine the effect of the variables in the data of learners and L1 English expert writers.

References

Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Edinburgh: Pearson Education Ltd.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. London: Longman.

Biber, Douglas and Susan Conrad. 2019. *Register, genre, and style*. Cambridge: Cambridge University Press.

Burnard, Lou. 2007. *Reference guide for the British National Corpus (XML Edition)*. URL: http://www.natcorp.ox.ac.uk/docs/URG/

Egbert, Jesse & Marianna Gracheva. 2022. Linguistic variation within registers: granularity in textual units and situational parameters. *Corpus Linguistics and Linguistic Theory* 19(1), 115-143.

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

Gries, Stefan Th. 2021. *Statistics for linguistics with R: A practical introduction*. Berlin, Boston: De Gruyter Mouton.

Halliday, Michael A. K. 1978. *Language as social semiotic: The social interpretation of language and meaning*. London: Arnold.
Larsson, Tove. 2019. Grammatical stance marking in student and expert production: Revisiting the informal-formal dichotomy. *Register Studies* 1(2), 243–268

Graphing registers: Exploring register differences via collocational networks in the BNC2014

Hanna Schmück (Lancaster University)

One of the many ways in which register differences and linguistic nuances within different fields of discourse can be explored and compared is via collocations, here broadly defined as a commonly co-occurring group or set of words (Barnbrook et al., 2013; Stulpinaitė et al., 2016). Previous research shows that register can be used to partially predict collocations in American English (Berber Sardinha, 2017), and underscores the significance of investigating how linguistic patterns contribute to the distinctive characteristics of various registers. This paper aims to explore register differences in the BNC2014 (v2; Love et al, 2017; Brezina et al., 2021) through the lens of collocational clusters derived from subcorpus-wide collocation networks Acknowledging the often-conflicting definitions of register and genre, this study adopts Biber and Conrad's (2019) interchangeable use of these terms, employing subcorpora of the BNC2014 as proxies for registers. The questions raised in this paper are how well collocation networks can capture register variation, which registers in the BNC2014 are the most collocationally similar/dissimilar, and what unique collocational clusters emerge from each subcorpus.

In order to address these, a multidisciplinary approach spanning corpus linguistics and graph theory is used to generate networks of *all* collocations* within each of the 8 subcorpora of the BNC: Academic Language, E-Language, Fiction, Magazines, Newspapers, Written-To-Be-Spoken, Official Documents, and Spoken Language. Following a methodology akin to Karaminis et al. (2023), the collocations are required to lie above both a $\Delta P_{forward}$ and logDice threshold to ensure robustness in terms of the forwards predictability and coherence of the collocations. The large linguistic networks (LLNs) generated on the basis of these subcorpora represent the collocational profiles, structure, and aboutness (Pecina, 2010; Xiao & McEnery, 2006; Brezina, 2016; Baker, 2016; Brezina et al., 2015) of the respective registers in the BNC2014. MCODE clustering (Bader & Hogue, 2003) is employed in order to identify the collocationally most closely-interconnected clusters unique to each register. Employing this methodology serves two purposes: firstly, it allows for a systematic and fully interpretable exploration of large-scale differences in collocational tendencies among different registers of modern British English, and, secondly, it showcases which new avenues for interpretation a novel approach to collocation visualisation can bring.

On a broader scale, notable collocational convergence is observed, with the highest overlap occurring between Magazines and News (20.2%), Spoken and Written-to-be-Spoken (19.9%), and E-Language and Spoken (19.6%). Conversely, the lowest overlap is evident between Official Documents and Written-to-be-Spoken (2.0%), Academic Language and Written-to-be-Spoken (2.3%), and Fiction and Official Documents (3.2%). Figures 1 and 2 show key clusters from the opposing domains of Official Documents and Written-To-Be-Spoken respectively. This novel multidisciplinary approach integrating corpus linguistics and graph theory with MCODE clustering presents a new way for systematically exploring register variation.

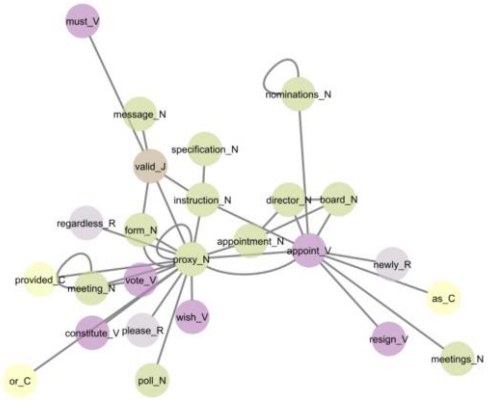


Figure 1. Subcluster containing collocates unique to Official Documents in the BNC2014

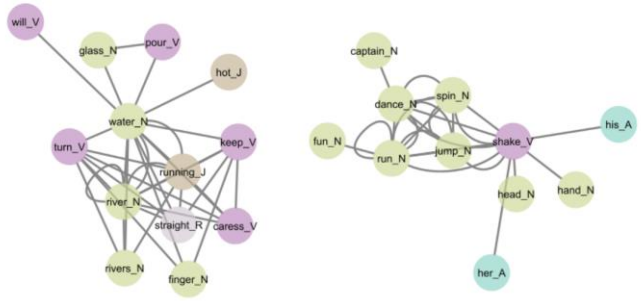


Figure 2. Subcluster containing collocates unique to Written-To-Be-Spoken in the BNC2014

*Note: CPN: POS-tagged lemma, ΔPforward and logDice, 90th percentile in both, sentence span.

References

Barnbrook, G., Mason, O. & Krishnamurthy, R. (2013). The concept of collocation. In G. Barnbrook, O. Mason & R. Krishnamurthy (eds.), *Collocation: Applications and Implications* (pp. 3–31). Palgrave Macmillan. https://doi.org/10.1057/9781137297242_1

Berber Sardinha, T. (2017). Lexical priming and register variation. In M. Pace-Sigge & K. J. Patterson (eds.) *Lexical Priming: Applications and Advances* (pp. 190–230). John Benjamins. https://doi.org/10.1075/scl.79.08ber

Biber, D. & Conrad, S. (2019). *Register, Genre, and Style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Bader, G. D. & Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4(2). https://doi.org/10.1186/1471-2105-4-2

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics* 21(2), 139–164. https://doi.org/10.1075/ijcl.21.2.01bak

Brezina, V. (2016). Collocation networks: Exploring associations in discourse. In P. Baker & J. Egbert (eds.), *Triangulating Methodological Approaches in Corpus Linguistic Research* (pp. 90–107). Routledge.

Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Brezina, V., Hawtin, A. & McEnery, T. (2021). The written British National Corpus 2014 – design and comparability. *Text & Talk* 41(5-6), 595–615. https://doi.org/10.1515/text-2020-0052

Karaminis, T., Gabrielatos, C., Maden-Weinberger, U. & Beattie, G. (2023). Portrayals of autism in the British press: A corpus-based study. *Autism* 27(4), 1092-1114. https://doi.org/10.1177/13623613221131752

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

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources & Evaluation* 44, 137–158. https://doi.org/10.1007/s10579-009-9101-4
Stulpinaitė, M., Horbačauskiene, J. & Kasperavičienė, R. (2016). Issues in translation of linguistic collocations: Lingvistinių kolokacijų vertimo ypatumai. *Studies about Languages* (29), 31–41. https://doi.org/10.5755/j01.sal.0.29.15056
Xiao, R. & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics* 27(1), 103–129. https://doi.org/10.1093/applin/ami045

On the spread of DO-support in Early Modern and Late Modern English

Ulrike Schneider (University of Frankfurt/University of Mainz)

The present study investigates the diachronic spread of DO-support in negated contexts. Around the time when DO-support was established (ca. 1400 – 1700), the English negation pattern gradually changed in negative declarative sentences from predominantly 'finite verb + *not*', as in (1), to '(finite) operator + *not* (+ lexical verb)', as in (2).

(1) I say not
(2) I do not say.

At the same time, operator DO also became common in questions as well as in negative imperatives (cf. e.g. Denison 1993: 451; Jespersen 1917: 10–11; Strang 1970: 151; Visser 1969: §1440–1441). The spread of DO did not proceed equally fast in all syntactic contexts, though. To this day, the go-to source for quantitative evidence of this phenomenon is Ellegård (1953). He shows, for instance, that the rate of DO-support in negative questions had already reached 75% around 1550, while it would take another 150 years for DO-support in negative declaratives to reach this rate. While Ellegård (1953) manually ploughed his way through nearly 300,000 sentences to assess whether they contained DO, the vast majority of these (over 95%) were affirmative declaratives. Thus, his rates of DO-support given for negative imperatives and negative questions are based on comparatively small datasets. Furthermore, his data does not cover the time after 1700. Data by Ecay (2015), who replicated Ellegård (1953), suggests that, by this time, the spread of DO-support was far from complete, particularly in negative imperatives.

The present study therefore starts out by replicating Ellegård's (1953) analyses of negated contexts based on data from large corpora. It utilises three corpora from the Chadwyck-Healey collection, i.e. Early English Prose Fiction, Eighteenth-Century Fiction and Nineteenth-Century Fiction, which represent British prose fiction published between 1500 and 1900. From these corpora, all instances of *not*, including the contractions *n't* and *'nt*, were extracted, resulting in a database of 800,000 tokens. All tokens where *not* co-occurs with a finite verb were then classified into negative imperatives, declaratives and questions of several types.

Besides assessing the spread of DO-support in negative contexts, the study also aims to determine the role that other operators, like modals and auxiliaries, played in this process as recent studies (cf. e.g. Budts 2022) suggest that they may have paved the way for operator DO.

References

Budts, Sara. 2022. A connectionist approach to analogy. On the modal meaning of periphrastic *do* in Early Modern English. *Corpus Linguistics and Linguistic Theory* 18(2), 337-364.
Denison, David. 1993. *English historical syntax*. London/New York: Longman.

Early English prose fiction. 1997–2015. Ed. by Holger Klein, David Margolies & Janet Todd. Chadwyck-Healey. ProQuest LLC.
Ecay, Aaron William. 2015. A multi-step analysis of the evolution of English *do*-support. PhD Thesis, University of Pennsylvania.
Eighteenth-century fiction. 1996–2015. Ed. by Judith Hawley, Tom Keymer & John Mullan. Chadwyck-Healey. ProQuest LLC.
Ellegård, Alvar. 1953. *The auxiliary do: The establishment and regulation of its use in English.* Stockholm: Almqvist and Wiksell.
Jespersen, Otto. 1917. *Negation in English and other languages*. Copenhagen: A.F. Høst.
Nineteenth-century fiction. 1999–2000. Ed. by Danny Karlin & Tom Keymer. Chadwyck-Healey. ProQuest LLC.
Strang, Barbara M. H. 1970. *A history of English.* London: Methuen.
Visser, Fredericus Theodorus. 1969. *An historical syntax of the English language – Syntactical units with two verbs. Vol. 3, first half*. Leiden: E.J. Brill.

Automated, corpus- and usage-based semantic classification of word classes

Martin Schweinberger and Chang-Hao Luo (University of Queensland)

The semantic classification of word classes, such as adjectives, nouns, and verbs, is integral to various linguistic analyses, including investigations into language variation and change (cf. e.g., Tagliamonte 2008; Schweinberger 2021), as well as language learning and teaching (e.g. Schweinberger 2020). The existing paradigm for semantic classification relies on predetermined categories (see, e.g. Dixon, 1977), often arbitrary, leading to inefficiencies and inconsistencies, particularly due to the manual allocation process. This study proposes an innovative approach by introducing an automated, usage-based semantic classification system utilizing word embeddings.

Word embeddings, fundamental to generative language models, capture the collocational profiles or co-occurrence patterns of words, thereby reflecting their contextual usage within language. By leveraging this technology, our approach provides a means to assess word similarity grounded in authentic language usage. The aim of this study is to present a method that alleviates the labor-intensive and non-reproducible nature of manual annotation processes, offering a more efficient solution for studies that rely on semantic word classifications.

These classifications are derived from word embeddings sourced from extensive British National Corpus (BNC, 2007) and the Corpus of Contemporary American English (COCA, see Davis, 2010), ensuring a diverse and comprehensive linguistic foundation. The high-dimensional word embeddings are subjected to dimension-reduction using t-SNE and the optimal number of categories (or classes) are determined using The result of this study is a corpus- and usage-based semantic classification of adjectives which are presented visually (see Figure 1 below) and a freely accessible list of adjectives, verbs, and nouns, and their semantic classification that can be used by the wider community. The use of this classification is exemplified in a case study that examines diachronic trends in the use of adjectives over time based on the Corpus of Historical American English (COHA, Davis, 2012). The results show that evaluative, emotional adjectives, and in particular positive emotional adjectives, have increased in frequency over time while neutral descriptive adjectives remain stable.

This paper not only introduces a novel methodology but also showcases how such classifications can be used in a study of adjective use across time and it aims to democratize access to automated, usage-based semantic classifications for adjectives, nouns, and verbs. This study not only introduces an innovative automated semantic classification system but also

contributes to the ongoing discourse on transparency and reproducibility in linguistic research. The openly accessible semantic classifications form a valuable resource for advancing studies that depend on precise, transparent, reproducible, and efficient semantic word classifications.

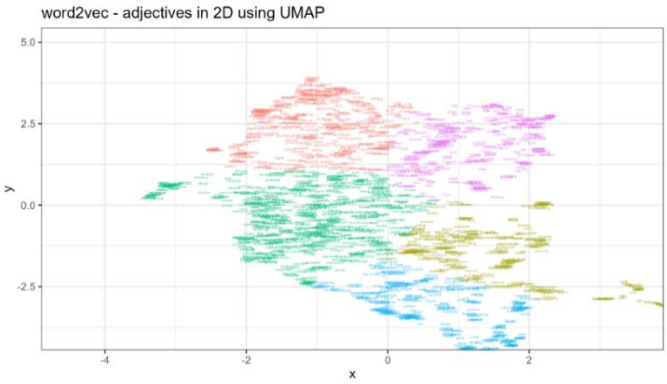


Figure 1. Visualisation of semantic adjective classes based on word embeddings based on the BNC.

References

BNC Consortium. (2007). *British National Corpus*. Oxford Text Archive Core Collection.

Davies, Mark. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25(4): 447-464.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora* 7(2): 121-157.

Dixon, R. M. W. (1977). Where have all the adjectives gone? *Studies in Language* 1: 19–80.

Schweinberger, Martin. (2020). A corpus-based analysis of adjective amplification among native speakers and learners of English. *International Journal of Learner Corpus Research* 6(2): 163-192.

Schweinberger, Martin. (2021). Ongoing change in the Australian English amplifier system. *Australian Journal of Linguistic* 41(2): 1-29.

Tagliamonte, Sali. (2008). So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics* 12(2): 361–394.

A corpus linguistic approach to anti-vaccination discourse in Victorian England

Elena Semino[1], Alice Deignan[2] and the Quo VaDis project team[1]

([1]Lancaster University, [2]University of Leeds)

Vaccine hesitancy is a highly topical phenomenon. Even before the Covid-19 pandemic, the World Health Organization described it as 'the reluctance or refusal to vaccinate despite the availability of vaccines', and included it among the top ten threats to global health, alongside antimicrobial resistance and climate change (WHO 2019). Anti-vaccination sentiment is not a modern phenomenon, however, but has existed for as long as vaccines have been available (Durbach 2005).

This talk is concerned with anti-vaccination discourse in Victorian England, following the introduction of compulsory vaccination against smallpox in the middle of the 19th century. After an overview of the historical background, we present the design and construction of the 3.6-million-word 'Victorian Anti-Vaccination Discourse Corpus' (VicVaDis). The corpus consists of 133 anti-vaccination pamphlets and other popular literature published in England between the 1853 Vaccination Act, which mandated smallpox vaccination for babies, and the 1907

Vaccination Act, which effectively ended the compulsory nature of vaccination. The corpus is freely available for researchers as a resource for the historical investigation of vaccine hesitancy.

In this talk we demonstrate how the corpus can be used to supplement historical studies in investigating Victorian anti-vaccination concerns and arguments. Specifically, we address the following research questions:

1) How can corpus linguistic methods be used to investigate the main anti-vaccination arguments in 19th century England?
2) How do these arguments compare with present-day arguments against vaccinations?

We employ concordance, collocation and keyness analyses to show how (compulsory) vaccination was presented as a threat to civil liberties, against nature, a distraction, and ineffective. We show how some of these concerns and arguments are specific to smallpox vaccination and to its mandatory nature at the time, for example with regard to skin infections following vaccination and fines for parents of unvaccinated children. However, the analysis also reveals many parallels with 21st-century vaccine hesitancy, including with regard to childhood vaccinations and vaccines against Covid-19. For example, there are parallels between the concerns about civil liberties in relation to compulsory smallpox vaccination and in relation to compulsory Covid-19 vaccination for travellers and for some professional groups during the 2020-2022 pandemic. Similarly, some Victorian writers claimed that smallpox could be prevented by increased cleanliness, in the same way as some present-day critics of HPV vaccination suggest that HPV infection can be prevented by sexual restraint (Hendry et al. 2013). We finish with some reflections on further avenues for the exploitation of the corpus.

References

Durbach, N. (2005) *Bodily Matters. The Anti-Vaccination Movement in England, 1853-1907*. Duke University Press, Durham NC.

Hendry, M., Lewis, R., Clements, A., Damery, S. & Wilkinson C. (2013) 'HPV? Never heard of it!': a systematic review of girls' and parents' information needs, views and preferences about human papillomavirus vaccination. *Vaccine* 25(45), 5152-5167.

World Health Organization. Top ten threats to global health in 2019. https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019.

A corpus phonological analysis of acoustic vowel variation in Pakistani English

Muhammad Shakir and Philipp Meer (University of Münster)

Work-In-Progress

Pakistani English (PakE) is the local L2 variety of English used in Pakistan. Most previous research on the variety has been on morphosyntax, lexis and code-switching, and register variation (e.g. A. Mahmood 2009; R. Mahmood 2009; Shakir & Deuber 2018, 2019; Shakir 2020). Although some local studies have analyzed select acoustic properties of PakE vowels (Bilal et al. 2011; Farooq & Mahmood 2017; Abbasi et al. 2018, Khan et al. 2020), a large-scale and systematic corpus phonological study of acoustic vowel variation in PakE is not available to date.

Addressing this research gap, the current study investigates acoustic vowel variation in PakE monophthongs and diphthongs using a corpus phonological approach. The overall aim is to provide a descriptive account of the Pakistani English vowel space and vowel inventory based on a sizable dataset. Specifically, we investigate the following research questions:

1) What is the overall shape of the acoustic vowel space in Pakistani English?
2) What is the general position of all lexical sets in the vowel space?
3) To what extent can differences in vowel inherent spectral change (VISC) be observed?
4) To what extent can differences in vowel quantity in terms of duration be found?

To that end, we build a larger dataset of PakE speakers using broadcast talks: n = 90. These programs have been aired between 2011-2023 on the public television broadcaster (Pakistan Television, PTV). Though the in-compilation International Corpus of English (ICE) Pakistan also includes some hand-annotated data in this genre, we apply automated speech recognition (ASR) and speaker diarization using WhisperX (Bain et al. 2023) to arrive at a larger corpus. The data is manually checked to correct speaker identification and any time overlap in transcriptions. At the time of writing, we have completed the manual correction of all 187 speakers.

Following ASR and manual correction, FAVE (Rosenfelder et al. 2014) is used for automatic segmentation, phonemic transcription, and automated (Bayesian) vowel formant (F1 and F2) estimation, which has been found to perform reliably on postcolonial Englishes (Meer 2020, Meer et al. 2021). F1 and F2 are estimated at five equidistant temporal locations of a vowel's duration (i.e. at 20, 35, 50, 65, and 80%). Following best practices, (i) all data are psychoacoustically transformed and normalized (Moore 2010: 459; Adank 2003), and (ii) vowels are analyzed with respect to both target-oriented measures of F1-F2 (monophthongs: at 50%; diphthongs: at 20 and 80%) and various dynamic acoustic aspects, including measures of spectral change and spectral rate of change (Jacewicz & Fox 2013; Farrington et al. 2018; Meer 2023). The acoustic data is analyzed using state-of-the-art statistical methods, including mixed-effects models. The results will provide the first large-scale corpus phonological account of variation in monophthongs and diphthongs in PakE.

References

Abbasi, Abdul, Mansoor Channa, Masood Memon, Stephen John, Irtaza Ahmed & Kamlesh Kumar 2018. Acoustic characteristics of Pakistani English vowel sounds. *International Journal of English Linguistics* 8(5): 27-34.

Adank, Patti. 2003. *Vowel normalization*: *A perceptual acoustic study of Dutch vowels*. Wageningen: Ponsen & Looijen.

Bain, Max, Jaesung Huh, Tengda Han & Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*.

Bilal, Hafiz Ahmad, Muhammad Asim Mahmood & Rana Muhammad Saleem 2011. Acoustic analysis of front vowels in Pakistani English. *International Journal of Academic Research* 3(6): 20–27.

Farooq, Mahwish & Asim Mahmood 2017a. Acoustic analysis of front vowels /Ɛ/ and /æ/ in Pakistani Punjabi English. *International Journal of English Linguistics* 8(1): 219. https://doi.org/10.5539/ijel.v8n1p219.

Farrington, Charlie, Tyler Kendall & Valerie Fridland. 2018. Vowel dynamics in the Southern Vowel Shift. *American Speech* 93(2): 186–222.

Jacewicz, Ewa & Robert A. Fox. 2013. Cross-dialectal differences in dynamic formant patterns in American English vowels. In Geoffrey S. Morrison & Peter F. Assmann (eds.), *Vowel inherent spectral change*, 177–198. Berlin: Springer.

Khan, Tariq, Farzana Masroor, Zulfiqar Ali & Naveed Nawaz Ahmad 2020. An acoustic analysis of back vowels among Punjabi English speakers in Pakistan. *Pakistan Journal of Languages and Translation Studies* 1: 169–190.

Mahmood, Muhammad Asim. 2009. *A corpus based analysis of Pakistani English*. Multan: Bahauddin Zakariya University PhD Thesis.

Mahmood, Rashid. 2009. *A lexico-grammatical study of noun phrase in Pakistani English*. Multan: Bahauddin Zakariya University PhD Thesis.

Meer, Philipp. 2020. Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. *The Journal of the Acoustical Society of America* 147(4): 2283-2294.

Meer, Philipp. 2023. *Standard English in Trinidadian secondary schools: Accent variation and attitudes* (PhD dissertation). University of Münster, Germany. https://doi.org/10.17879/18988575787

Meer, Philipp, Brato, T. & J. A. Matute Flores. 2021. Extending automatic vowel formant extraction to New Englishes: A comparison of different methods. *English World-Wide* 42(1): 54-84.
Moore, Brian C. 2010. Aspects of auditory processing related to speech perception. In William J. Hardcastle, John Laver & Fiona E. Gibbon (eds.), *The handbook of phonetic sciences*, 2nd edn., 454–488. Hoboken, NJ: Wiley-Blackwell.
Rosenfelder, I., J. Fruehwald, K. Evanini, S. Seyfarth, K. Gorman, H. Prichard & J. Yuan. 2014. FAVE (Forced Alignment and Vowel Extraction). Program Suite v1.2.2. https://github.com/JoFrhwld/FAVE.
Shakir, Muhammad & Dagmar Deuber 2018. A multidimensional study of interactive registers in Pakistani and US English. *World Englishes* 37(4): 607–623. https://doi.org/10.1111/weng.12352.
Shakir, Muhammad & Dagmar Deuber 2019. A multidimensional analysis of Pakistani and U.S. English blogs and columns. *English World-Wide* 40(1): 1–23. https://doi.org/10.1075/eww.00020.sha.
Shakir, Muhammad 2020. *A corpus based comparison of variation in online registers of Pakistani English using MD analysis*. Münster: University of Münster PhD Thesis.

Metadiscourse patterns in discourse variation: A comparative corpus-based study of research articles in life sciences

Lilia Shevyrdyaeva (Shenzhen MSU-BIT University)

Disciplinary academic writing as a form of knowledge construction undergoes continuous change over time reflecting processes both inside and outside of academia. Each disciplinary community establishes and shares genre conventions and pragmatic strategies reflecting, to a certain degree, the scientific research it conducts. Modern day's pressure to publish increases the value of metadiscourse markers as effective tools of making a paper accepted by the disciplinary community. Metadiscourse markers contribute to building a convincing argument by structuring a text, projecting the author's standpoint, engaging the audience, establishing credibility, etc. (Hyland, 2005).

Previous studies have observed variation in patterns of metadiscourse markers both between and within disciplines (Gillaerts & Van de Velde, 2010; McGrath & Kuteeva, 2012; Cao & Hu, 2014; Hyland & Jiang, 2018). Drawing on Hyland's framework (Hyland, 2018), this paper examines how academic authors with different disciplinary expertise use metadiscourse markers in the introduction and discussion sections of their research writing to mark the epistemic stance and establish a relationship with their audiences. This paper compares three closely related disciplines representative of the genre conventions, narrative tradition and language use in the life sciences – Ecology, Genetics and Immunology – to describe the variation of metadiscourse patterns. To this end, three sub-corpora were compiled of research articles from top-ranking disciplinary journals – 85 papers each (more than 150,000 words) – published in 2019-2021 and authored by L1 English speakers. Both interactive and interactional metadiscourse markers were analyzed in combination with statistical methods and correlation analysis.

This comparative corpus-based investigation describes the frequency and distribution of metadiscourse markers across the sections of research papers and identifies specific patterns characteristic of each sub-genre. Quantitative and qualitative analyses reveal inter-disciplinary variation and similarities between three academic discursive traditions. The most informative interactional markers exhibiting distinct differences turned out to be self-mention and, predictably, hedges and boosters, whereas for interactive markers interesting correlations were observed for code glosses, evidentials and transition markers, particularly, in introductions.

Such comprehensive approach to metadiscourse markers provides a convenient tool for the description of cross-disciplinary genre variation in addition to pragmatic strategies and allows to hypothesize on the driving forces behind such patterns. The findings of this study can be used in genre-based ESAP writing instruction and inform genre analyses across academic disciplinary discourses.

References

Cao, F. & Hu, G. (2014). Interactive metadiscourse in research articles: A comparative study of paradigmatic and disciplinary influences. *Journal of Pragmatics 66*, 15–31. https://doi.org/10.1016/j.pragma.2014.02.007

Gillaerts, P. & Van de Velde, F. (2010). Interactional metadiscourse in research article abstracts. *Journal of English for Academic Purposes* 9(2), 128–139. https://doi.org/10.1016/j.jeap.2010.02.004

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7(2), 173–192. https://doi.org/10.1177/1461445605050365

Hyland, K. (2018). *Metadiscourse: Exploring Interaction in Writing* (Bloomsbury Classics in Linguistics). Bloomsbury Academic.

McGrath, L. & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes* 31(3), 161–173. https://doi.org/10.1016/j.esp.2011.11.002

Hyland, K. & Jiang, F. K. (2018). 'In this paper we suggest': Changing patterns of disciplinary metadiscourse. *English for Specific Purposes* 51, 18–30. https://doi.org/10.1016/j.esp.2018.02.001

Going forward as an adverbial: Register and the spread of innovation

Olli Silvennoinen (University of Helsinki)

Temporal adverbials (or adjuncts) very seldom take the form of non-finite clauses (Hasselgård 2010: 38). When a participial clause does have a temporal function, it typically refers to an event that either precedes that of the main clause or occurs at the same time (Fonteyn & Van De Pol 2016: 195–196). This presentation will examine a pattern that flies in the face of both of these generalisations: the metaphorical use of *going forward* as a future adverbial roughly synonymous with 'in the future' or 'from now on', illustrated in (1)-(3):

(1) He said his focus **going forward** will be on jobs and the economy for the district and all Virginia. (COHA, 2017, NEWS, Virginian-Pilot)
(2) **Going forward**, the team will look to gel quickly, as the team's next three PDL matchups are all on the road. (COHA, 2017, NEWS, Arizona Daily Star)
(3) We were expecting to run into some issues here **going forward**. (COHA, 2017, NEWS, Washington Times)

Existing accounts of future expressions in English have not discussed *going forward* (e.g. Bergs 2010). According to the OED, futurate *going forward* originated in the 1970s in American English, and its primary domains of use are business and management (OED, s.v. *go forward* 2). Formally, this use of *going forward* is an *ing*-clause used as a free adjunct (Kortmann 1991). Discursively, *going forward* often projects an optimistic view of the future where the subject of the clause it modifies tends to be in control of events (cf. Bhatia 2008), as in (1), although this kind of reading is not obligatory (as shown in (3)).

In this presentation, the use and development of futurate *going forward* will be charted from the perspective of register variation and semantic change. The hypothesis is that the

development of *going forward* is characterised by growing functional and stylistic versatility, which should be shown in its appearance in a broader set of genres as well as grammatical and semantic contexts. The data comes from the Corpus of Historical American English (COHA) and the Corpus of Contemporary American English (COCA), and it will be analysed from the perspectives of form, function and register distribution.

According to preliminary results, futurate *going forward* first appears in COHA in the 1990s. After two decades of relatively modest rates of occurrence, its frequency suddenly increases in the 2010s. In terms of register, it is particularly common in newspapers and magazines, but in the 2010s, it also spreads to fiction and TV and film scripts. These findings are in line with the hypothesis of increasing versatility, which indicates a growing degree of conventionalisation in the language.

References

Bergs, Alexander. 2010. Expressions of futurity in contemporary English: A Construction Grammar perspective. *English Language and Linguistics* 14(2). 217–238. https://doi.org/10.1017/S1360674310000067.

Bhatia, Vijay K. 2008. Towards critical genre analysis. In Vijay K. Bhatia, John Flowerdew & Rodney H. Jones (eds.), *Advances in discourse studies*, 166–177. London & New York: Routledge.

Fonteyn, Lauren & Nikki Van De Pol. 2016. Divide and conquer: The formation and functional dynamics of the Modern English *ing*-clause network. *English Language and Linguistics* 20(2). 185–219. https://doi.org/10.1017/S1360674315000258.

Hasselgård, Hilde. 2010. *Adjunct adverbials in English* (Studies in English Language). Cambridge: Cambridge University Press.

Kortmann, Bernd. 1991. *Free adjuncts and absolutes in English: Problems of control and interpretation* (Germanic Linguistics). London & New York: Routledge.

Oxford English Dictionary Online. Oxford: Oxford University Press. www.oed.com

---

Competitive research funding discourse: Move structure, stance and engagement in successful European project summaries

Jolanta Šinkūnienė (University of Vilnius)

Over the past few years there has been a renewed surge of interest in science communication patterns, particularly in genres related to competitive research funding discourse. Some studies have focused on communication patterns, lexico-grammatical features, rhetorical-discursive conventions in the websites of international projects funded by such global EU research programs as H2020 (Corona, 2021; Lafuente-Millán, 2023; Mur-Dueñas, 2023). The results of these studies are very useful as they outline the specific practices of how research results can be presented to different audiences and stakeholders in order to increase their visibility. There is also a growing body of research which looks into the discourse of successful grant proposals for international competitive funding (Cotos, 2019; Matzler, 2021, Millar et al., 2022; Millar et al. 2023). The results of these studies show that researchers employ increasingly more promotional "hype" language (Millar et al. 2022) and that there is evidence that the writing style can influence the success of the grant proposal (see, for example, van den Besselaar & Mom 2022). Such studies are interesting and important as they analyse internationally successful and skillfully written texts from the best research teams worldwide, however, the number of such studies is still quite scarce.

The focus of this paper is on the summaries of European Cooperation in Science and Technology (COST) Actions provided on the COST digital platform and the summaries of

European Research Council (ERC) projects funded under the Consolidator grant scheme. COST Actions are networking projects, whereas ERC grants finance research projects. The summaries of both types of projects provide essential information about the project. The focus of this corpus-based qualitative and quantitative analysis is on 100 summaries of COST Actions which were granted COST funding in 2020-2022, and 100 summaries of ERC Consolidator grant winning projects funded in the same time period. Following Swales (1990, 2004) and Swales & Feak (2010), the rhetorical structure of each description is analysed together with the distribution of stance and engagement markers (following Hyland 2005) in different moves and steps in order to answer the research question whether the type of the project (networking vs research) influences the level of promotionalism in its description.

The preliminary results show that the rhetorical structure in both types of project summaries is quite promotional, combining elements from research article abstracts and CARS model of research article introductions. However, much more promotional discourse can be observed in research project summaries in comparison to networking project summaries. In both, however, stance markers prevail over engagement markers which suggests that in this genre it is more important for the researchers to highlight their stance rather than engage with the reader. The results of this research could be useful to potential applicants of European research or networking grants, or scholars interested in the construction of promotional hype discourse in English.

References

Corona, I. (2021) A window to the world: visual design and research visibility of European research projects' homepages. *European Journal of English Studies* 25(3), 352-368.

COST (European Cooperation in Science and Technology): https://www.cost.eu

Cotos, E. (2019) Articulating societal benefits in grant proposals: Move analysis of broader impacts. *English for Specific Purposes* 54, 15-34.

European Research Council (2023) ERC at a glance. In *About ERC*. Online document. 15 August 2023 https://erc.europa.eu/.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7(2), 173–192.

Lafuente-Millán, E. (2023). Evaluation and genre evolution in H2020 European research project websites: A multiple perspective of analysis. In O. Dontcheva-Navratilova, C. Pérez-Llantada, M. Bondi and J. Schmiedt (eds.), *Patterns of Language Variation and Change in Academic Writing: A Corpus-Based Diachronic Perspective*. Special Issue of *Token*, in press.

Matzler, P. P. (2021) Grant proposal abstracts in science and engineering: A prototypical move-structure pattern and its variations. *Journal of English for Academic Purposes* 49, 100938.

Millar, N., Batalo, B. and Budgell, B. (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, N., Mathis, B., Batalo, B. and Budgell, B. (2023) Trends in the expression of epistemic stance in NIH research funding applications: 1985-2020. *Applied Linguistics* XX, 1-18. https://doi.org/10.1093/applin/amad050

Mur-Dueñas, P. (2023). Exploring researchers' professional digital discursive practices: A genre analysis of European research project websites. *Ibérica 45*, 79-107.

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

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

Swales, J.M., & Feak, Ch. (2010). From text to task: Putting research on abstracts to work. In M.F. Ruiz-Garrido, J.C. Palmer-Silveira and I. Fortanet Gómez (eds.). *English for Professional and Academic Purposes*. Amsterdam/New York: Brill Rodopi. 167–180.

van den Besselaar, P. and Mom, C. (2022). The effect of writing style on success in grant applications. *Journal of Infometrics* 16(1), 101257.

Disciplinary variation in the expression of stance: A corpus analysis of research articles in hard and soft sciences

Elizaveta Smirnova and Javier Pérez-Guerra (HSE University, University of Vigo)

The linguistic representation of stance in academic discourse, which is defined as 'the expression of the speaker's or writer's personal feelings, attitudes, value judgements, or assessments' (Biber et al., 1999: 966), has been a widely-researched topic in the field of academic discourse studies (see, for example, Hyland, 1999; Biber, 2006; Salager-Meyer et al., 2012). However, the number of studies investigating disciplinary variation in the expression of stance in professional academic writing remains relatively scarce. This investigation undertakes a quantitative analysis of stance features, carried out on a 1,597,000-word corpus of research papers in four 'soft' sciences (business studies, linguistics, history, and political science) and four 'hard' sciences (mathematics, engineering, chemistry, and physics), published in leading peer-reviewed journals. The goal is twofold: first, to describe the markers of stance employed by professional authors and, second, to test the hypothesis that there is significant variation in the expression of stance across disciplines.

Since the recognition of the morphological encoding of subjectivity in English is controversial (Aikhenvald, 2006; Boye & Harder, 2009), this investigation relies on the exploration of lexical and grammatical strategies denoting subjectivity, taken from previous research on stance in academic discourse, mainly Biber's (2006) and Hyland's (2005) lists of stance expressions. The lists comprise modal and semi-modal verbs, stance adverbs, complement clauses of different types, hedges, boosters, attitude markers and expressions of self-mention. First, the corpus was processed with AntConc (Anthony, 2014), which eased the retrieval of the features. Second, the frequencies of the stance markers and the disciplines were modelled statistically in an attempt to both weigh the contribution of the features to the hard/soft variation, and to determine similarities/differences among the hard and the soft disciplines as far as stance is concerned.

The analysis of the data revealed not only significant differences in the use of the stance markers between the hard and the soft sciences but also the validity of stance expressions as proxies for disciplinary categorisation at least in scientific discourse. For instance, it was found that *that*-complement clauses controlled by verbs tend to occur more frequently in the hard-science papers, specifically in the results and literature review sections. Expressions of self-mention also turned out to be more commonly employed by the hard sciences. This finding is at odds with Hyland's (2005: 181) conclusion that 'in the sciences it is common for writers to downplay their personal role to highlight the phenomena under study'.

The results are expected to aid in the development of genre-specific language resources for EAP (English for Academic Purposes) courses. Understanding how different disciplines express stance can help learners to write more effectively in their respective fields. Besides, the analysis of disciplinary variation in the expression of stance may contribute to the development of automated tools for text evaluation that can assist researchers, editors and reviewers in identifying the discipline-specific stance features that are pervasive or expected in the discourse of research articles.

References

Aikhenvald, A. Y. (2006). *Evidentiality*. Oxford University Press.

Anthony, L. (2014). AntConc (Version 3.4.4) [Computer Software]. Tokyo: Waseda University.

Biber, D. (2006). *University language: A corpus-based study of spoken and written Registers.* John Benjamins.

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

Boye, K. & Harder, P. 2009. Evidentiality: Linguistic categories and grammaticalization. *Functions of Language* 12, 65–86.

Hyland, K. (1999). Disciplinary discourses: Writer stance in research articles. In C. N. Candlin, & K. Hyland (eds.) *Writing: Texts, processes and practices* (pp. 99–121). Routledge.
Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse studies* 7(2), 173–192.
Salager-Meyer, F., Ariza, M. Á. A., & Briceño, M. L. (2012). The voice of scholarly dispute in medical book reviews, 1890–2010. In K. Hyland & C.S. Guinda (eds.) *Stance and voice in written academic genres* (pp. 232–248). Palgrave Macmillan.

Lexicons of prophetic women

Jeremy Smith (University of Glasgow)

Early 1650s English society had been turned upside down by civil war and the king's beheading: 'It was a hinge in the world's history. God was about to do something new' (Ryrie 2017: 118). An outpouring of print resulted, reflecting a remarkable efflorescence of religious/socio-political innovation. In particular, the war had 'released women into the public world of contention, and into speech and writing' (Hobby 2001: 174). Such women included the 'Fifth Monarchist' millenarian Anna Trapnel (fl. 1642-1660), who used prophecy to mount a comprehensive attack on contemporary institutions, or the Quaker Mary Howgill (c.1620-?1666), who denounced Oliver Cromwell – to his face – as 'a stinking dunghill in the sight of God' (Cromwell listened to her, and let her go). Both women exemplify two communities of practice, with distinctive behaviours, beliefs, and 'ways of talking'.

This paper's primary research question is: is it possible – in a statistically robust and comprehensive way – to identify such 'ways of talking'? Contemporaries certainly thought so. Lucy Hutchinson (1620-1689) describes her earnestly Calvinist husband thus (1904: 125):

> … the godly of those dayes, when he embrac'd their party, would not allow him to be religious because his hayre was not in their cutt, nor his words in their phraze, nor such little formallities altogether fitted to their humor …

'Phrazes', it seems, developed special theological meanings depending on confessional orientation. Hutchinson may have been thinking of such passages as the following, from the Quaker leader Margaret Fell's (1614-1702) *A Testimonie of the Touchstone* (1656: 4):

> … if ever ye come to know the living God, ye must turn your minds to the light which is in you, which Christ Jesus saith, take heed that the light that is in you be not darkness; for then how great is that darkness. And so all your blinde guides which keeps your minds from this light, which is in you. So your minds being from it, it is to you darkness, and so then how great is that darkness. But turning your minds to this light, and joining your mindes to it, and hearkening to it; then will ye come to see this blinde eye opened.

We might note 'phrazes' such as 'turning your minds to this light', a collocation Fell frequently uses to align with Quaker notions of the 'light within' (see Roads 2018), and other words in the same semantic field (*dark*, *see*, *blind*, *eye*).

In this paper, part of a larger project on the English religious lexicon's historical evolution deriving from earlier preliminary studies (Smith 2020, 2021), two curated corpora of 1650s prophetic women's writings – Fifth Monarchist and Quaker – are examined. These corpora are compared with a contemporary large reference corpus, supplied from EEBO-TCP using *Semantic*

EEBO. Toolkits used to establish salient lexical patterns include Laurence Anthony's *AntConc*, Lancaster UCREL's *Log-Likelihood and Effect Size Calculator*, and Sheffield's *Linguistic DNA Concept Modeller*. Other resources harnessed are the *Oxford English Dictionary* and the *Historical Thesaurus of English*. Specialised lexicons thus identified are contextualised, contributing to the developing field of theolinguistics (see e.g. Hobbs 2021).

This presentation is supported by a grant from the Leverhulme Trust.

References

AntConc: https://www.laurenceanthony.net/software/antconc/
Early English Books Online (EEBO) – Text Creation Partnership: https://quod.lib.umich.edu/e/eebogroup/
Historical Thesaurus of English: https://historicalthesaurus.arts.gla.ac.uk/
Linguistic DNA: https://www.linguisticdna.org/
Log-Likelihood and Effect Size Calculator: https://ucrel.lancs.ac.uk/llwizard.html
Oxford English Dictionary: https://www.oed.com/
Semantic EEBO: https://www.english-corpora.org/eebo/

Fell, Margaret 1656. *A Testimonie of the Touchstone*. London: Simmons.
Hobbs, Valerie 2021. *An Introduction to Religious Language*. London: Bloomsbury.
Hobby, Elaine 2001. Prophecy, enthusiasm and female pamphleteers, in N.H. Keeble (ed.) *The Cambridge Companion to Writing of the English Revolution*. Cambridge: University Press, 162-178.
Hutchinson, Lucy (ed. Harold Child) 1904. *Memoirs of the Life of Colonel Hutchinson*. London: Kegan Paul, Trench and Trübner.
Roads, Judith 2018. *The Distinctiveness of Quaker Prose 1650-1699*. PhD Birmingham.
Ryrie, Alec 2017. *Protestants: The Radicals who Made the Modern World*. London: Collins.
Smith, Jeremy J. 2020. Godly vocabulary in Early Modern English religious debate, in Eva Jonsson and Tove Larsson (eds.) *Voices Past and Present - Studies of Involved, Speech-Related and Spoken Texts, in Honor of Merja Kytö*. Studies in corpus linguistics 97. Amsterdam: John Benjamins, 96-112.
Smith, Jeremy J. 2021. Lexical choices in Early Modern English devotional prose. *Journal of Historical Pragmatics* 22 (special issue in honour of Andreas Jucker), 264-282.

Sensitivity of dispersion measures to distributional patterns and corpus design

Lukas Sönning and Jesse Egbert (University of Bamberg, Northern Arizona University)

The purpose of dispersion measures is to quantify how widely or evenly an item (or structure) is distributed in a corpus (see Gries 2008, 2020). Recent work, however, has shown that these measures also respond to other features in the data. Thus, *D* varies systematically with the number of texts (or corpus parts) that enter the analysis (Biber et al. 2016), and virtually all measures also vary systematically with the frequency of an item (Gries 2022). The present study aims to provide further insights into the sensitivity of dispersion measures to various aspects of corpus design and data distribution; overall, we consider six evenness measures (*D, D2, S, DP, DA, DKL*) and text dispersion (*TD*) as a pervasiveness index.

In line with recommendations issued in earlier work (Burch et al. 2017; Egbert et al. 2020), we measure dispersion across linguistically meaningful units: the text files in the corpus. Building on Biber et al. (2016) and Gries (2022), our sensitivity analysis considers the number of texts (*n*) and the frequency of an item (*f*). We further take into account two aspects of corpus design that are relevant for text-level analyses: the average text length (*l*) and the variability of text lengths (*v*). Finally, we also consider the feature of interest: the dispersion of an item (*d*).

Our general approach is to vary these factors across realistic values, to observe how dispersion measures respond to changes in the data. To mimic frequency distributions observed in actual corpus data, we use the negative binomial distribution, which has been applied successfully to word frequency distributions (e.g. Mosteller and Wallace 1964; Church and Gale 1995). To obtain realistic and representative scenarios, the parameter values we implement are trained using data from the Brown Corpus. Our simulation study uses a full factorial design with n ∈ {50, 200, 1000 texts}, f ∈ {0.1, 1, 10 ptw}, l ∈ {500, 2000, 10000 words}, v ∈ {constant, bell-shaped, flat/rectangular}, and d ∈ {1, .60, .30, .10}.

Our results suggest that, within the settings covered by our analysis, frequency accounts for the greatest share of variation in estimates, between roughly 20% (*D*) and 50% (*TD*). Our manipulation of dispersion, on the other hand, explains between 20% (*D, TD*) and 30% (D_A, D_{KL}) of the variability. The mean text length also reverberates in dispersion estimates, accounting for roughly 10% (*D, D2*) to 15% (*TD, DP*) of the observed variation. We also note that *D2* is similar to *D* in that it increases systematically with the number of texts in the corpus, making it unsuitable for text-level dispersion analyses. Our findings therefore support Gries's (2022) pessimistic view on the behavior of dispersion measures, even though some measures appear to be more responsive to the actual feature of interest. In addition, we observe that dispersion measures are also affected by the average text length, which may compromise their utility for genre (or corpus) comparisons.

References

Biber, D., R. Reppen, E. Schnur & R. Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4), 439–464.

Burch, B., J. Egbert & D. Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2), 189–216.

Church, K. W. & W. A. Gale. 1995. Poisson mixtures. *Natural Language Engineering* 1(2), 163–190.

Egbert, J., B. Burch & D. Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1), 89–115.

Gries, S. Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4), 403–437.

Gries, S. Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 99–118. New York: Springer.

Gries, S. Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2), 171–205.

Mosteller, F. & D. L. Wallace. 1984. *Applied Bayesian inference: The case of The Federalist Papers.* New York: Springer.

Regression and random forests: Synergies for variationist corpus research

Lukas Sönning, Jason Grafmiller and Raquel P. Romasanta

(University of Bamberg, University of Birmingham, University of Santiago de Compostela)

Two widely used modeling strategies in current corpus-based work are regression- and tree-based methods (see, e.g., Bernaisch 2022; Grafmiller 2023). Both paradigms offer certain advantages for variationist analysis: While a regression approach allows researchers to build theoretically grounded models that can embrace variation at multiple levels in the data (e.g. Winter 2020; Tizón-Couto & Lorenz 2021), tree-based procedures excel in convenience and flexibility (e.g. Strobl et al. 2009; Levshina 2021): they require few(er) data-analytic decisions, make few(er) distributional assumptions, and are able to capture complex relationships in the

data. Most studies currently tend to focus on one form of analysis, without giving full consideration to the value of complementary strategies. The aim of this paper is to show how analytical leverage may be gained by combining the strengths of both methods, the goal being to arrive at a descriptively adequate model of the data, with appropriate indications of statistical uncertainty.

We present a case study dealing with the alternating usage of *that*- and -*ing*-complement clauses (CCs) after the verb *regret* (see Romasanta 2023). The data are drawn from the British and American component of GloWbE (Davies 2015) and include 1,112 tokens and six predictor variables: subject coreferentiality, temporal relation between main clause and CC, negation/voice/length of CC, and the presence of intervening material between *regret* and the CC. This illustrative setting may be considered quite typical of much variationist work, as it involves a combination of internal linguistic features that are hypothesized to be associated with CC variation.

The confirmatory elements in our analysis incline us toward a regression framework, which allows us to adapt our model to the intended scope of inference (e.g. by including random effects). With little guidance in the literature on functional relationships and interactions between predictors, our working model has a simple fixed-effects structure. As a safeguard against oversimplifying patterns in the data, we exploit the flexibility of tree-based methods. To this end, we make use of predictive margins (see Sönning & Grafmiller 2023), which allow us to query a random forest for interactions and non-linear relationships. This allows us to note whether elaborations to our (over)simplified regression structure are needed. This dialog between the models is directed by our research priorities, background knowledge, and the statistical uncertainty surrounding the random forest predictions. We demonstrate how convergence between the analysis modes may strengthen our confidence in the results, and demonstrate how to strategically look for indications of interactions and non-linearities that may remain masked using a main-effects-only, straight-line regression. We also draw attention to some limitations of this concerted effort, including the fact that categorical regression models operate on a non-linear link scale (e.g. log odds in the case of logistic regression), which may compromise the comparability of interaction patterns (see Loftus 1978).

References

Bernaisch, T. 2022. Comparing generalised linear mixed effects models, generalised linear mixed-effects model trees and random forests. In O. Schützler & J. Schlüter (eds.), *Data and methods in corpus linguistics: Comparative approaches*, 163–193. Cambridge: Cambridge University Press.

Davies, M. 2015. Introducing the 1.9 billion word Global Web-Based English Corpus (GloWbE). *The 21st Century Text* 5.

Grafmiller, J. 2023. Visualizing grammatical similarities in comparative variationist analysis. In L. Sönning & O. Schützler (eds.), *Data visualization in corpus linguistics: Critical reflections and future directions*. Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/22/grafmiller/

Levshina, N. 2021. Conditional inference trees and random forests. In M. Paquot & S. Th. Gries (eds.), *A practical handbook of corpus linguistics*, 611-643. New York: Springer.

Loftus, G. R. 1978. On interpretation of interactions. *Memory & Cognition* 6(3), 312-319.

Romasanta, R. P. 2023. *Probabilistic variability in clausal verb complementation in World Englishes.* Frankfurt: Peter Lang.

Sönning, L. & J. Grafmiller. 2023. Seeing the wood for the trees: Predictive margins for random forests. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2022-0083

Strobl, C., J. Malley & G. Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4), 323–348.

Tizón-Couto, D. & D. Lorenz. 2021. Variables are valuable: Making a case for deductive modeling. *Linguistics* 59(5), 1279-1309.

Winter, B. 2020. *Statistics for linguists: An introduction using R*. New York: Routledge.

That was real(ly) nice: A corpus-based study on particularities of intensifying *real* and *really*

Ulrike Stange-Hundsdörfer (University of Mainz)

While intensifiers in general have received ample attention in the past couple of decades (see, for instance, the studies by Tagliamonte 2008 and Tagliamonte and colleagues 2002, 2003, 2005, or by Aijmer 2018a, b, 2020 and Stenström 1999, 2000), the contrast between *really* and the zero form *real* as intensifiers is typically mentioned in passing only (see Opdahl 2000 on variation on other dual-form adverbs). What we know so far extends basically to two aspects: first, across different L1 varieties, the intensifier *real* tends to be associated with male speech (Tagliamonte & Ito 2002, Yaguchi et al. 2010, D'Arcy 2015) and/or teenage speech (Stenström et al. 2002; Aijmer 2018a), and second, *real* is perceived to be of lower formality compared to *really* (Tagliamonte & Ito 2002; Aijmer 2018a).

The present study systematically explores the use of intensifying *real* and *really*, focusing on linguistic factors (e.g., syntactic position, collocational patterns, syllable structure of modified item) as well as speaker-related variables (age and gender). It considers their use with adjectives in attributive (1) and predicative position (2) and with adverbs (3).

(1) You're right. I guess we do have a *really* big problem. (DAYS 2009)
(2) She seems *real* upset, being stood up like that. (DAYS 2003)
(3) He's gonna be here with us *real* soon. (DAYS 2008)

The analysis draws on four samples consisting of c. 250 occurrences each for *real* ADJ, *really* ADJ, *really* ADV and c. 20 occurrences for *real* ADV (fewer occ. in the dataset than for the other items) that were extracted from the *Corpus of American Soap Operas* (2001-2012; Davies 2011-; 12.7m words for *Days of Our Lives*). In the samples, male characters showed by far the highest usage of *real* (no age effect), while younger female characters produced the highest number of tokens for the intensifier *really*. Still, for both genders the frequency of use for *really* ADJ increased in the time span 2001-2012, while *real* ADJ experienced a notable decline in frequency of use.

Real(*ly*) ADV remained stable as regards the distribution relative to one another but increased slightly in overall frequency. Interestingly, in bigrams of the type *real* ADV, adverbs with the suffix *-ly* are next to non-existent. The default form is that of the zero form (e.g., *Let me see that real quick*. – DAYS 2009; $p<0.001$***). While syllable structure did not affect the distribution of *real* vs. *really* with adjectives, it did significantly so with adverbs ($p<0.001$***; see also Rohdenburg & Schlüter 2009).

As regards *real*(*ly*) ADJ, *really* was frequently repeated (e.g., *I'm just … really, really sad*. – DAYS 2011), whereas in utterances with *real*, it was often the adjective that was copied (*She's a nice girl. Real nice*. – DAYS 2006; $p<0.001$***).

Differences were also detected concerning the construction in which *real*(*ly*) ADJ occurred. In the sample, *really* ADJ occurred significantly more frequently in predicative position than *real* ADJ ($p<0.001$***). This use has been associated with sounding affective and listener-oriented (Yaguchi et al. 2010). *Real* and *really* thus appear to not only pattern differently in the data but to also perform different functions.

References

Aijmer, Karin. 2018a. *That's well bad*. Some new intensifiers in spoken British English. In: Vraclav Brezina, Robbie Love & Karin Aijmer (eds.), *Corpus approaches to contemporary British English*, 60-95. New York: Routledge.

Aijmer, Karin 2018b. Intensification with *very*, *really* and *so* in selected varieties of English. In Sebastian Hoffmann, Andrea Sand, Sabine Arndt-Lappe & Lisa-Marie Dillmann (eds.), *Corpora and lexis* 106-139. Leiden: Brill Rodopi.

Aijmer, Karin. 2020. *That's absolutely fine*: An investigation of 'absolutely' in the spoken BNC2014. In Paula Rautionaho, Arja Nurmi & Juhani Klemola (eds.), *Corpora and the changing society*, 143-161. Amsterdam: Benjamins.
D'Arcy, Alexandra. 2015. Stability, stasis and change: The longue durée of intensification. *DIA* 32 (4): 449-493.
Davies, Mark. 2011-. Corpus of American Soap Operas. Available online at https://www.english-corpora.org/soap/.
Ito, Rika & Sali Tagliamonte. 2003. *Well* weird, *right* dodgy, *very* strange, *really* cool: Layering and recycling in English intensifiers. *Language in Society* 32: 257-279.
Opdahl, Lise. 2002. *LY or zero suffix? A study on variation of dual-form adverbs in present-day English*. Volumes 1 & 2. Frankfurt a.M.: Peter Lang.
Rohdenburg, Günther & Julia Schlüter. 2009. New departures. In: Rohdenburg, Günther & Julia Schlüter (eds.), *One language, two grammars*, 364-423. Cambridge: Cambridge UP.
Stenström, Anna-Brita. 1999. *He was really gormless – She's bloody crap*. Girls, boys and intensifiers. In: Hilde Hasselgard & Signe Oksefjell (eds.), *Out of corpora. Studies in honour of Stig Johansson*, 69-78. Amsterdam: Rodopi.
Stenström, Anna-Brita. 2000. *It's funny enough, man.* Intensifiers in teenage talk. In: John M. Kirk (ed.) *Corpora galore. Analyses and techniques in describing English*, 177-190. Amsterdam: Rodopi.
Stenström, Anna-Brita, Gisle Andersen & Ingrid K. Hasund. 2002. *Trends in teenage talk. Corpus compilation, analysis and findings*. Amsterdam: Benjamins.
Tagliamonte, Sali. 2008. *So different and pretty cool!* Recycling intensifiers in Toronto, Canada. English *Language and Linguistics* 12 (2): 361-394.
Tagliamonte, Sali & Rika Ito. 2002. Think *really different*: Continuity and specialization in English dual form adverbs. *Journal of Sociolinguistics* 6 (2): 236-266.
Tagliamonte, Sali & Chris Roberts. 2005. So weird; so cool; so innovative: The use of intensifiers in the television series Friends. *American Speech* 80 (3): 280-300.
Yaguchi, Michiko, Yoko Iyeiri & Yasumasa Baba. 2010. Speech style and gender distinctions in the use of very and real/really: An analysis of the Corpus of Spoken Professional American English. *Journal of Pragmatics* 42: 585–597.

The role of speaker attitudes in the consolidation of Gibraltar English in the twenty-first century

Cristina Suárez-Gómez (University of the Balearic Islands)

Most Gibraltarians in the 20th century were multilingual, using English as the institutional language along with two other, coexisting languages, Spanish and Llanito, these commonly used for non-institutional communication. However, in the 21st century English has become dominant, especially among younger speakers, suggesting that Gibraltar is steadily moving towards monolingualism, with globalisation and Brexit playing significant roles in this process. Together with language use, the attitudes of speakers towards a language constitute a crucial factor in the consolidation of that language within a territory, especially in World Englishes (Schneider 2007: 49; Bernaisch and Koch 2015: 119; see also Extra- and Intra-Territorial Forces Model, EIF, Buschfeld and Kautzsch 2020: 4).

The aim of this presentation is to analyse the current sociolinguistic landscape of Gibraltar by focussing on the attitudes shown by Gibraltarians towards the coexisting languages in the territory (English, Spanish and Llanito) in different domains. This analysis is based on responses to an anonymous questionnaire designed for this purpose and distributed online in 2021 (43 participants) and in person in 2023 (73 participants). It consists of 19 questions and, together with questions relating to participants' personal information, includes questions covering the

following issues: (i) their mother tongue and their parents' mother tongue; (ii) language use in different contextual domains; and (iii) attitudes towards these languages. Crucially, the results of the questionnaire are complemented by the results of an analysis of the manifestation of attitudes through various forms of expression (e.g. lexical sequences which contain the terms English, Spanish and Llanito, references to language, etc.; cf. Graedler 2014) in different registers of written texts from the Gibraltarian component of the ICE corpus (ICE-GBR).

A preliminary analysis confirms that the shift towards English-speaking homes began several decades ago, reinforced by the fact that the L1 of most of the participants' parents is English. Despite this shift towards monolingualism, the Gibraltarian community continues to self-identify as bilingual (or multilingual), not only the oldest generation (over 71 years), most of these being ESL speakers, but also the youngest generation included in the analysis (15-20 years), most of these being ENL speakers who no longer speak Spanish and/or Llanito. The results also reveal that Gibraltarians take pride in their multilingualism, while at the same time they prefer the exonormative British English accent over their Gibraltarian English one.

In the case of Gibraltar, these preliminary results reflect the complex and hybrid identity of Gibraltarian speakers as already observed in previous qualitative studies (Seoane 2016; Sanchez 2017). In general, the study confirms the role of extra- and intra-territorial forces in shaping language attitudes and identity construction (cf. EIF model, Buschfeld and Kautzsch 2020), which are themselves key in the development of a language variety, Gibraltar English in this case.

References

Bernaisch, Tobias and Christopher Koch 2015. Attitudes towards Englishes in India. *World Englishes* 35(1): 118–132.

Buschfeld, Sarah and Alexander Kautzsch 2020. Introduction. In Sarah Buschfeld and Alexander Kautzsch (eds.), *Modelling World Englishes: A Joint Approach to Postcolonial and Non-Postcolonial Varieties*. Edinburgh: Edinburgh University Press, pp. 1–15.

Graedler, Anne-Linne. 2014. Attitudes towards English in Norway: A corpus-based study of attitudinal expressions in newspaper discourse. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication* 33: 291-312

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

Seoane, Elena 2016. Telling the true Gibraltarian story: An interview with Gibraltarian writer M.G. Sanchez. *Alicante Journal of English Studies* 29: 251–258.

Sanchez, M. G. 2017. Representing Gibraltarianness. *Gibraltar Chronicle*. www.chronicle.gi/representing-gibraltarianness. Last accessed: 15 July 2021.

The discourse markers *hello* in English and *hallo* in Norwegian informal spoken registers: A cross-linguistic, corpus-based investigation

Jan Svennevig and Ingrid Kristine Hasund (University of Agder)

The present study is a cross-linguistic, corpus-based investigation of the discourse markers *hello* in English and *hallo* in Norwegian. As discourse markers, *hello/hallo* are both used with non-vocative functions primarily associated with informal spoken registers. 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 that has explored the use of *hallo* as a discourse marker in modern Norwegian. Using data from three corpora of informal spoken Norwegian – 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 two main functions.

The first is as a reproach to an addressee for having said or done something inappropriate, incorrect, irrelevant or illegitimate (henceforth, the *reproach function*). The second is to preface the announcement of something especially newsworthy, interesting, or important (henceforth, the *announcement function*). The reproach is responsive to some previous action, whereas the announcement is projecting what is to come next. According to Svennevig, there is reason to believe that the reproach function is a pragmatic borrowing from English *hello;* the origin of the announcement function is unclear. The findings are supported by Andersen (2014) in a study of the influence of English on Norwegian pragmatics, where it is suggested that the reproach function is a pragmatic borrowing from English. Andersen does not mention the announcement function in his study but calls for more comparative studies that explore the range discourse functions of markers such as *hello/hallo.*

The present study aims to fill this research gap by comparing the findings from Svennevig's (2012) study to data from English to explore to what extent *hello* and *hallo* are used with similar discourse functions and possibly to explore the emergence of the respective functions. As the Norwegian corpora were compiled in the 1990s and early 2000s, the English data will be drawn from the spoken parts of the British National Corpus, BNC1994, the CallHome corpus and the SBCSAE corpus. The methodological approach used is collection-based Conversation Analysis, whereby instances are excerpted from the corpora and analyzed qualitatively in their interactional and sequential context to discern distinct usages (Sidnell, 2013). Our research questions are: Which discourse functions does *hello* have in English? To what extent do they correspond to the functions of Norwegian *hallo*? Preliminary results seem to indicate a great deal of overlap.

References

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

Sidnell, J. (2013). Basic conversation analytic methods. In T. Stivers & J. Sidnell (Eds.), *The Handbook of Conversation Analysis* (pp. 77–99). John Wiley & Sons.

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

The beginnings of the genitive alternation in Old and early Middle English

Roxanne Taylor, Tine Breban and Kersti Börjars

(University of Huddersfield, The University of Manchester, University of Oxford)

The genitive alternation, exemplified in (1), is a well-known feature of Present-Day English (PDE).

(1) the mayor's son
the son of the mayor

Contrary to the assumption that the alternation found in PDE developed in the late Middle English (ME) period, once *of* came to express 'pure possession' (Rosenbach 2002), we argued (Taylor et al. 2022), using data from *York-Helsinki-Toronto Parsed Corpus Old English Prose* (YCOE, Taylor et al. 2003), that there is evidence for a genitive alternation in Old English (OE), involving genitive morphology and phrases with *of*, that could plausibly stand in direct continuity to the alternation in PDE. Variation between genitive case and phrases with *of* in OE exists in the expression of a limited range of semantic relations, most of which, such as part-whole relations, are excluded from the genitive alternation in later stages of English (see also Mitchell 1985; Allen

2008; Anderson 2016; Ceolin 2021). However, some argument relations appear to be found both in the OE variation and the PDE alternation, e.g. (2).

(2) dysegra manna herunga
unwise.GEN.PL man.GEN.PL praise
'unwise men's praise' (*cocathom1,ÆCHom_I,_28:416.169.5563*)

idle herunge of mannum
idle praise of man.DAT.PL
'idle praise of men' (*coverhom, HomS_38_[ScraggVerc_20]:145.G.2631*)

Prepositions other than *of* alternated with genitive morphology as well in argument relations in OE.

The aim of this paper is to bridge the gap between the findings for OE and the corpus studies of the alternation in late ME (Rosenbach and Vezzosi 2000, Rosenbach et al. 2000, Rosenbach 2002). Our study uses data from the *Penn Parsed Corpus of Middle English* (PPCME2, Kroch et al. 2000), and complements and expands on Allen (2008), which shows that the *s*-genitive stops being used to mark part-whole relations and certain objective argument relations in early ME, thereby removing these relations from the envelope of variation. We searched the 1150-1250 period of the PPCME2 initially for nouns and pronouns marked as possessors, and then for those same possessum head nouns with prepositional phrases. Finally we searched for all *of*-phrases. We found, in line with earlier studies (Thomas 1931), that the use of *of* increases overall. Yet, as in OE, other prepositions are used for relations also marked by the *s*-genitive, including *towart,* which did not participate in the alternation in YCOE. The use of *of* as marker of argument relations increases, primarily in the marking of themes and stimuli. We confirm that there is indeed alternation that can be traced from OE in certain argument relations, notably agent ones, but also some others. New semantic relations are added to the envelope of variation as *of* is used to mark a wider range of relations in variation with the *s*-genitive, including body-part relations, as in (3).

(3) þe wunden of ure Lauerdes flesch
the wounds of our lord-GEN flesh
'the wounds of our Lord's flesh' (*CMANCRIW_1,II.215.3109*)

Þeose heorte wunden
these heart-GEN wounds
'these wounds of the heart' (*CMANCRIW_1,II.201.2883*)

We conclude that the starting point of the PDE genitive alternation is not the increase of *s*-genitive into the territory of the *of*-genitive in late ME, but rather the establishment of the *of*-genitive in OE and its expansion in early ME.

References

Allen, Cynthia L. 2008. *Genitives in early English: Typology and evidence*. Oxford: Oxford University Press.
Anderson, Salena Sampson. 2013. Genitive variation in Old English verse with special attention to *Beowulf*. *English Studies* 94(8), 845–871.
Ceolin, Andrea. 2021. Constraints on Old English genitive variation. *Journal of Historical Syntax: Proceedings of the 20th Diachronic Generative Syntax (DiGS) conference* 5, 1–35.
Kroch, Anthony, Ann Taylor & Beatrice Santorini. 2000-. *The Penn-Helsinki parsed corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania. https://www.ling.upenn.edu/ppche/ppche-release-2016/PPCME2-RELEASE-4.
Mitchell, Bruce. 1985. Old English Syntax. Vol 1. Oxford: Clarendon Press.

Rosenbach, Anette. 2002. *Genitive variation in English: Conceptual factors in synchronic and diachronic studies*. Berlin and New York: Mouton de Gruyter.
Rosenbach, Anette, Dieter Stein & Letizia Vezzosi. 2000. On the history of the *s*-genitive. In Ricardo Bermúdez-Otero, David Denison, Richard Hogg & C. B. McCully (eds.), *Generative theory and corpus studies: A dialogue from 10ICEHL*, 183–210. Berlin and New York: Mouton de Gruyter.
Rosenbach, Anette & Letizia Vezzosi. 2000. Genitive constructions in Early Modern English: New evidence from a corpus analysis. In Rosanna Sornicola, Erich Poppe and Ariel Shisha-Halevy (eds.), *Stability, variation and change of word-order patterns over time*, 285–307. Amsterdam and Philadelphia: John Benjamins.
Taylor, Ann, Anthony Warner, Susan Pintzuk & Frank Beths. 2003. *The York–Toronto–Helsinki parsed corpus of Old English prose*. http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm.
Taylor, Roxanne, Tine Breban & Kersti Börjars. 2022. The genitive alternation in early English. Paper presented at ICHL 25, University of Oxford. 1-5 August 2022.
Thomas, Russell. 1931. *Syntactic processes involved in the development of the adnominal periphrastic genitive in the English language*: University of Michigan dissertation.

The Oxford Corpus of Historical English: Developing a new resource for linguistic and other academic research in modern and historical English

Amanda Thomas, Claire Etty and Kate Wild

(Oxford English Dictionary (Oxford University Press))

We are in the early stages of planning a new, large, high-quality corpus of historical English, with a dual purpose: to facilitate work on editing the Oxford English Dictionary, and – in collaboration with partners in the wider academic community – to develop a key new resource for scholarly research.

Currently, OED editors have access to a range of historical corpora, which together provide a patchwork representing some periods of English, from various geographical areas. There is a need for more extensive and consistent coverage across the full historical and geographical span. Our contemporary corpora include the Oxford English Corpus, and a large corpus of online news articles with wide geographical coverage. Our existing historical corpora draw on sources such as Early English Books Online, the Corpus of Historical American English, and a collection of books held by the Bodleian Libraries and digitized by Google Books. There are limitations in the time periods, regions, and text-types represented in these current resources. The proposed new corpus would dramatically increase the scope and quality of both historical and contemporary corpora available to editors.

The project has an ambitious scope, bringing together into a single resource English language data from 1500 to the present day, covering as many varieties of World English in as much chronological depth as possible. This will vastly increase the diversity of high-quality corpus data available to lexicographers and other academics working with historical English data, including in the fields of linguistics, history, literature, and social sciences, enabling ground-breaking and innovative research projects.

The new corpus project has identified historical sources to meet criteria for historical and regional coverage, with additional factors such as genre, domain and register also informing sourcing decisions. We are actively seeking to partner with external experts, collaborating to ensure a high level of data quality to meet a wide range of research use cases.

The large scale of the project brings technical challenges which we expect to approach with a combination of traditional and novel natural language processing methods, including machine

learning and large language models. Areas being actively considered include lemmatization, PoS tagging, semantic tagging at the word level, and automatic domain recognition at the text level. The methods chosen to perform these NLP tasks will be selected for and adapted to the different periods and regions of the corpus's content.

A prototype of the new historical corpus will be created initially for internal use, with the breadth of the content to be built out in the following years. Eventually, we expect the corpus to be available alongside the OED.com platform, allowing people querying the dictionary to access additional corpus-based information, and vice versa, for a richer user experience. We anticipate that the project will be transformative in enabling new research, as well as improving the accuracy and value of OED editorial work.

Detecting, analysing and visualizing semantic change: Collocational divergence in English and Czech

Ondřej Tichý and Václav Cvrček (Charles University)

With the recent advances in the areas of word embeddings, transformers and large language models, quantitative approaches to semantics and semantic change have received increasing attention (Laicher et al. 2021). While these approaches are undeniably successful in many respects, the interpretation of their results in terms of more traditional systematic historical linguistics remains difficult (Tahmasebi et al. 2021).

In this paper, we propose an innovative data-driven approach to detecting, analysing and visualizing semantic change based on comparison of collocational profiles. One of the key elements of our methodology is the use of the Kullback-Leibler divergence (DKL), a measure of how one probability distribution (entropy) diverges from another. In our context, DKL quantifies the extent of change in the distribution of collocates over different time periods. We can readily interpret our results as the level of surprise caused by the change in the context of the same lexical item over time. This method enables us to systematically represent and quantify semantic change, a task traditionally challenging in linguistics.

To enhance the accessibility of our research, we have introduced a set of interactive tools (DIACOL) integrated into the Czech National Corpus (CNC) infrastructure allowing users to query, analyse, and visualize changes in diachronic corpora hosted by the CNC infrastructure.

To test our approach, we have applied DIACOL to data from the Corpus of Historical American English (COHA) and the SYN v. 11 corpus of Czech. Initially, we looked at well-known or expected examples of semantic change in the two respective periods (19^{th}-20^{th} century for English and 1996-2021 for Czech) to establish a baseline of divergence that signal potential semantic change. The level of the baseline turned out to be arbitrary to a degree and also not universal – it may differ especially with respect to diachronic distance of the samples, their overall consistency etc.

With the initial thresholds set, we have created randomized samples of several thousand lexical items from various frequency bands. In these samples, we have been able to detect a number of easy-to-interpret examples. In English, *plane* has broadened from a term of geometry to also mean airplane after the invention of heavier than air flight, in Czech *rouška* ("veil", "face mask") has shifted significantly with the COVID pandemic. Other examples were more surprising, like *disorder* (civic > mental) or *flour*. DIACOL allows us to analyse these potential candidates further by looking at the collocates that contribute most to the collocational divergence at both ends of the diachronic gap. In a number of examples (like English *flour*) and especially over short diachronic spans (like Czech *plyn, "natural gas",* in 2020s), the divergence is largely due to either

the change in the composition of the dataset (the boom of internet recipes in the case of *flour*) or the rise of a popular, but possibly transitory topic (the war in the Ukraine in case of *plyn*).

While it is one of the remaining challenges of our methodology to better tease out such changes in usage from changes more traditionally characterized as semantic shifts, our approach seems to showcase how the latter originates in the former.

References

Davies, M. (2010) *The Corpus of Historical American English (COHA)*. Available online at https://www.english-corpora.org/coha/.

Hnátková, M., Křen, M., Procházka, P., Skoumalová, H. (2014). The SYN-series corpora of written Czech. In *Proceedings of the* Ninth *International Conference on Language Resources and Evaluation (LREC'14)*, 160–164. Reykjavík: ELRA. ISBN 978-2-9517408-8-4.

Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2021). Explaining and improving BERT performance on lexical semantic change detection. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 192–202.

Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y. & Hengchen, S. (eds.) (2021). *Computational approaches to semantic change*. Language Science Press. https://langsci-press.org/catalog/book/303

Grammaticalization, reduction and the emergence of variants: The *sort/kind/type of X* construction in spoken American English

David Tizón-Couto and David Lorenz (University of Vigo, Lund University)

This paper presents a new approach to the English construction *sort/kind/type of X* (SKT). Previous research has documented its grammaticalization from binomial (N *of* N) to qualifying, adverbial and pragmatic marker (cf. Ajmer 1984, Brems & Davidse 2010, Margerie 2010, Denison 2011, Reichelt 2021; consider the function in *I like this kind of music* vs *I kind of like this music*). Desemanticization, decategorialization and phonetic reduction have been frequently discussed in connection with the SKT; however, only one study has focused on a phonetic aspect (prosody), and found that increasing grammaticalization is associated with decreasing prosodic prominence (Dehé & Stathi 2016).

The present study investigates the relation between grammaticalization and phonetic reduction in the SKT construction synchronically. If grammaticalization affects phonetic form, we would expect that more grammaticalized usage types come with more reduction (erosion). On the other hand, reduction can also result from articulatory factors (speaking rate, phonological context), social context or item frequency. By way of a detailed analysis of actual realizations in a large data set, we can pitch these factors of reduction against each other to test whether grammaticalization really has a reducing effect in spontaneous usage. Moreover, we can map out the pronunciation variants of SKT items, and relations between them, in American English, while most previous studies had focused on British English.

We present a quantitative analysis of SKT items extracted from two different corpora of North American spoken English: the Santa Barbara Corpus of Spoken American English (Du Bois et al. 2000-2005), of spoken conversation, and the Buckeye Corpus (Pitt et al. 2007), of personal interviews. While the comprehensive analysis is still in progress, we can observe differences from previous results on British English (e.g. Dehé & Stahi 2016), such as a higher number of *kind of* sequences and more adverbial uses of the construction. Especially *kind of* appears to show an overall pattern of reduction that partly confirms the hypothesis that more

grammaticalized forms are more backgrounded and reduced. However, elision of the final fricative (as in "kinda") is compromised by phonological environment (such as a following vowel), suggesting that there are hurdles to the entrenchment of *kinda/sorta* as distinct grammaticalized variants.

References

Ajmer, Karin. 1984. *Sort of* and *kind of* in English conversation. *Studia Linguistica* 38: 118–128.

Brems, Lieselotte & Kristin Davidse. 2010. The grammaticalisation of nominal type noun constructions with *kind/sort of*: chronology and paths of change. *English Studies* 91 (2): 180– 202.

Dehé, Nicole & Katerina Stathi. 2016. Grammaticalization and prosody: The case of English *sort/kind/type of* constructions. *Language* 92(4): 911–946.

Denison, David. 2011. The construction of SKT. Paper presented at the 2nd Vigo-Newcastle-Santiago-Leuven International Workshop on the Structure of the Noun Phrase in English (NP2), 15–16 September 2011. Online: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:172513

Du Bois, John W., Robert Engelbertson, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson & Nii Martey. 2000-2005. *Santa Barbara Corpus of Spoken American English*, Parts 1-4.

Pitt, Mark A., Laura C. Dilley, Keith Johnson, Scott Kiesling, William D. Raymond, Elizabeth Hume & Eric Fosler-Lussier. 2007. *Buckeye Corpus of Conversational Speech (2nd release)* [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Reichelt, Susan. 2021. Recent developments of the pragmatic markers *kind of* and *sort of* in spoken British English. *English Language and Linguistics* 25(3): 563–580.

The grammar of fake news: Corpus linguistics meets machine learning

Zia Uddin[1], Nele Põldvere[2] and Aleena Thomas[1]

([1]SINTEF Digital, [2]University of Oslo)

In this study, we investigate the grammar of fake news by bringing together insights from corpus linguistics and machine learning. While the former offers a robust corpus-based register analysis of grammatical features, namely, multidimensional analysis (Biber, 1988), the latter contributes with methodological capabilities for the automatic identification of fake news based on the features. Fake news detection has made remarkable progress in natural language processing and machine learning (e.g., Rashkin et al., 2017; Põldvere et al., 2023), but it has not taken full advantage of the linguistic resources that are available. Based on the new PolitiFact-Oslo Corpus (Põldvere et al., 2023), we aim (i) to describe the grammatical differences between fake and real news across a variety of text types in a large corpus, and (ii) to develop a deep learning-based efficient approach for fake news detection based on these differences.

A common distinction in multidimensional register analysis is between informational and involved styles of communication. While the former tends to contain more nouns and is common in registers with dense styles of communication such as news reportage, the latter is characterized by a more frequent use of pronouns, verbs and adjectives and is common in spontaneous conversation with lower levels of information density. Departing from the view that fake news is a register in its own right, Grieve and Woodfield (2023) analyzed 49 grammatical features in a small collection of fake and real news texts by one journalist. They found fake news to be more similar to involved styles of communication through the use of, e.g., present tense verbs, emphatics and predicative adjectives. This was different from real news which shared features with informational styles of communication.

In contrast to Grieve and Woodfield (2023), in this study we make use of a large corpus of fake and real news in English: the PolitiFact-Oslo Corpus. The main strengths of the corpus are that the texts have been individually labelled for veracity by experts and are accompanied by important metadata about the text types (e.g., social media, news and blog) and sources (e.g., X, *The Gateway Pundit*). At present, the corpus contains 428,917 words of fake and real news, and it is growing. To extract the grammatical features, we used the Multidimensional Analysis Tagger (Nini, 2019), followed by a deep learning-based efficient approach (Attention-based Long Short-Term Memory; LSTM) to train the features incriminating fake and real news. The trained model was then used to automatically detect the fake news texts.

The preliminary results based on a sample from the corpus indicate that there are systematic differences between fake and real news, which by and large are indicative of the distinction between involved and informational styles of communication, respectively. However, these differences are not the same across the text types, with social media showing lower levels of information density in fake news than news and blog. Our machine learning model based on the grammatical features also shows promising results (LSTM mean accuracy: 90%), particularly when compared to models without the grammatical features.

References

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

Grieve, J. & Woodfield, H. (2023). *The language of fake news.* Elements in Forensic Linguistics. Cambridge University Press. https://doi.org/10.1017/9781009349161

Nini, A. (2019). The Multi-dimensional analysis tagger. In T. Berber Sardinha & M. Veirano Tinto (eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 67–94). Bloomsbury Academic.

Põldvere, N., Uddin, Z. & Thomas, A. (2023). The PolitiFact-Oslo Corpus: A new dataset for fake news analysis and detection. *Information* 14, 627. https:// doi.org/10.3390/info14120627

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S. & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In M. Palmer, R. Hwa & S. Riedel (eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931–2937). Association for Computational Linguistics.

From manual to automatic annotation of co-speech gesture in multimodal corpora

Peter Uhrig[1], Ilya Burenko[2], Irina Pavlova[3] and Anna Wilson[3]

([1]FAU Erlangen-Nürnberg, [2]TU Dresden, [3]University of Oxford)

In many domains of linguistic research, the manual annotation of data is a prerequisite to the theoretical analysis of regularities and variation. While the automatic annotation of text has come a long way (at least in English) and many types of annotation (e.g. part of speech, lemma) are only rarely done manually, the automatic annotation of the audio signal and in particular the video signal is still comparably limited, with most groups working on co-speech gesture annotating data manually in great detail and with very high accuracy, but at a low speed.

In this paper, we describe the process of the creation of a dataset that contains both detailed manual annotations on selected snippets and a set of automatically-generated annotations on the full collection. The data is taken from the show *SophieCo Visionaires*, which ran on RT (formerly *Russia Today*) and which we study in the context of a project on disinformation. The videos and their subtitles were downloaded from YouTube, turned into a corpus (see Dykes/Wilson/Uhrig 2023 for technical details) and made available via CQPweb (Hardie 2012).

The manual annotation process provided us with an understanding of what features of co-speech gesture we wanted to annotate automatically. We developed the automatic annotation in a constructive and continuous dialogue between the linguistic analysis and the computer vision analysis, gradually expanding and improving the latter.

The first set of automatic video annotations included whether the show host is present, is speaking, and whether her hands are visible in the image (see Uhrig et al. 2023 for details). These annotations were directly included in the CQPweb corpus and can thus be queried. For the development of further automatic annotations such as gesture zones and gesture directions, a different process was needed to allow for immediate visual inspection and verification. Here we relied on ELAN, which allowed us to display the video stream and all annotations together. We leveraged deep learning technologies such as OpenPose (Cao et al. 2016) and built a set of rule-based tools that processed OpenPose's body pose keypoints. For gesture direction, we relied on smoothed and normalized keypoint coordinates in the lateral and vertical dimensions, approximated by horizontal and vertical coordinates in each video frame. For gesture zones, we worked with a dynamic, speaker-based grid similar to that of McNeill's division of the gesture space (1992: 89), which we again normalized to the speaker's size on the screen.

While many current methods of multimodal corpus linguistics rely on laboratory recordings with controlled settings, our automatic approach works in the wild, with scene changes and various camera perspectives as well as multiple speakers on screen.

In the final part of the presentation, we will explain the benefits but also the limitations of automatic gesture annotation using examples from our dataset.

References

Cao, Zhe, Gines Hildago, Tomas Simon, Shih-En Wei & Yaser Sheikh. 2019. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43(1), 172-186.

Dykes, Nathan, Anna Wilson & Peter Uhrig. 2023. A pipeline for the creation of multimodal corpora from YouTube videos. *Proceedings of KONVENS 2023*.

ELAN (Version 6.7) [Computer software]. 2023. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Hardie, Andrew. 2012. CQPweb: Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3), 380–409.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

Uhrig, Peter, Elinor Payne, Irina Pavlova, Ilya Burenko, Nathan Dykes, Mary Baltazani, Evie Burrows, Scott Hale, Philip Torr & Anna Wilson. 2023. Studying time conceptualisation via speech, prosody, and hand gesture: Interweaving manual and computational methods of analysis. In W. Pouw, J. Trujillo, H. R. Bosker, L. Drijvers, M. Hoetjes, J. Holler, L. Van Maastricht, E. Mamus & A. Ozyurek (eds.), *Gesture and Speech in Interaction (GeSpIn) Conference*. doi:10.17617/2.3527220.

What-what reduplication in South African English: Anything but a question

Bertus van Rooy and Roné Wierenga (University of Amsterdam, University of Gent)

The reduplication of the interrogative *what-what* is a recent new usage in South African English that appears in online comments from the mid-2000s, and also in the NOW corpus from 2011. Of the 77 instances of *what-what* in NOW, only 15 are not from South Africa. Moreover, the non-South African usages are always interrogatives, as in (1), either in emphatic usage or imitations of the spoken language. The South African usages are only occasionally interrogative, but mostly perform other grammatical roles, such as noun or modifying adjective in (2) and (3) respectively.

(1) I'd seen up close the sharp edge of the sub judice rule. The sub what-what? Yes, it's Latin for ('before the court')… (NOW, Canada, 2018)
(2) TM1 used organs of state to screw Zuma, got caught and never got to finish his second term. No three strikes what-what. Out. Don't come Monday. (NOW, South Africa, 2016)
(3) If it is not women's awards or miss what-what pageants, it has to be dramatic local artists… (NOW, South Africa, 2016)

No other interrogative adverb or pronoun is used in reduplicated form in South African data in the TVSA or NOW corpora, with less than a dozen instances of all other interrogatives together in the data from the remaining 19 countries represented in NOW.

This paper analyses the use of *what-what* reduplications in the TVSA corpus of online comments (2005-2015) and the NOW corpus (2011-present) with a view to determining their semantic and syntactic characteristics. The noun usage is dominant, while adjective and discourse marker usages also occur freely throughout. All the syntactic options are used from the earliest time of attestation, but there is some change over time in the semantic prosody, with a negative sentiment becoming more dominant after an initial more even distribution of neutral and negative sentiments. One possible source of transfer was the possibility that this is syntactic borrowing from the ancestral languages of the South African Indian speech community, as illustrated by (4), that got generalised to the rest of South African English (Mesthrie, 1992). However, the borrowed usage, which is indeed attested in Indian South African usage, is still interrogative.

(4) What-what she told me I listened nicely. (Mesthrie, 1992: 204)

Thus, a broader search for answers and origins is required, including possible borrowing from vernacular Cape Afrikaans, where a few rare instances have been discovered going back to the late 1980s, but no consistent record of usage either, and limited precedent in the semantics of Afrikaans reduplication. A number of key events supporting diffusion, such as the use in the title of fiction (*Portrait with Keys: Joburg & what-what*, Vladislavic 2006) and two very popular television advertisements from 2008 and 2011 contributed to the uptake of the construction.

References

Mesthrie, Rajend. 1992. *English in Language Shift*. Johannesburg: Wits University Press.
Vladislavić, Ivan. 2006. *Portrait with Keys: Joburg & what-what*. Johannesburg: Umuzi.

Chunking and reanalysis at the individual level

Svetlana Vetchinnikova (University of Helsinki)

Usage-based constructionist approaches maintain that linguistic structure emerges from language use, notably through the process of grammaticalization. For example, the future marker *be going to* was originally an instance of a more general construction Sbj *be* Verb-ing *to* Verb expressing 'purpose' together with other possible verbs such as *journeing, traveling and returning* (Bybee 2006; Danchev & Kyto 1994). With a growth in frequency, it became entrenched as a chunk, underwent reanalysis, lost the association with the more general schema, acquired a new pragmatic function, and finally reduced to *gonna.* Does the same process take place at the individual level, i.e. within one's personal language use?

As my data, I will use 10 longitudinal corpora of comments posted on a single blog by different individuals, native and non-native speakers of English, over 8 years. The largest individual corpus comprises 1.75 million words while the rest vary between 250 and 40 thousand words. The comments of over 4 thousand occasional commenters (ca. 3.5 million words in total) on the same blog serve as a reference corpus representing the communal level. In a case study (Vetchinnikova forthcoming), I used the alternation between the full and the reduced form of *it is (it's)* as a diagnostic of reanalysis in the largest individual corpus. First, I categorized a total of 10,000 corpus occurrences of *it is/it's* into 14 frequent constructions. Then, I identified lexical items filling the open slot in all constructions and used delta P statistic to compute the degree to which a lexical item associates with a construction and the degree to which a construction associates with a lexical item (Gries & Ellis 2015). I also calculated normalized entropy for each construction as a measure of dispersion (Gries & Ellis 2015; Gries 2021). Finally, I built a logistic regression model predicting the reduced form as an effect of time, constructional entropy, and unidirectional word-to-construction and construction-to-word associations. Lexically specified instantiations of constructions were included as random effects. The model showed significant main effects of constructional entropy and construction- to-word associations confirming that reduction is a viable diagnostic of chunkedness in a corpus of a single person's language use. In addition, different lexical instantiations showed substantial variation in the extent to which they associate (intercept variance = 0.39) or become associated (slope variance = 0.17) with the reduced form over time suggesting idiosyncratic reanalysis as a function of usage. In fact, the model with random effects explained 33% (conditional R2) of variance and only 22% with fixed effects only (marginal R2).

This paper conducts the same analysis with the rest of the individual corpora and the communal corpus to test the extent to which the phenomenon of idiosyncratic reanalysis generalizes across individuals. In addition, it compares the effects in individual and communal corpora and in native and non-native speaker usage.

References

Bybee, Joan L. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language* 82(4), 711-733. https://doi.org/lü.1353/lan.2006.0186.

Danchev, Andrei & Merja Kyto. 1994. The construction be going to + infinitive in Early Modern English. In Dieter Kastovsky (ed.), *Studies in Early Modern English,* 59-78. De Gruyter. https://doi.org/10.1515/9783110879599.59.

Gries, Stefan Th. 2021. *Statistics far Linguistics with R: A Practica! Introduction. Statistics far Linguistics with R.* De Gruyter Mouton. https://doi.org/10.1515/9783110718256.

Gries, Stefan Th. & Nick C. Ellis. 2015. Statistical measures for usage-based linguistics. *Language Learning* 65(Sl), 228-255. https://doi.org/10.1111/lang.12119.

'Modal + *be going to*' and 'modal + *be about to*': An analysis in terms of grammaticalization and temporal structure

Naoaki Wada (University of Tsukuba)

This paper explains the differences in the distribution pattern of "modal + *be going to*" (BGT) and "modal + *be about to*" (BAT) in terms of degree of grammaticalization and their temporal structures. To my knowledge, no studies have analyzed them in detail. BGT has been analyzed in numerous studies from various perspectives (e.g. Brisard 2001; Eckardt 2006; Gesuato and Facchinetti 2011; Langacker 1990; Nicolle 1997; Tagliamonte, et al. 2014), while BAT has been considered only in some studies (e.g. Höche 2010; Mee 2013; Watanabe 2011). However, few

studies have compared the two forms systematically in a general theory of tense. Wada (2000, 2019) is one such study, so I take it as my explanatory basis.

His claim is mainly twofold: (i) unlike BGT, BAT is not a frozen unit and thus less grammaticalized because it can cooccur with other copula-like verbs (**seem/appear going to* vs. *seem/appear about to*) or *as if* (**as if going to* vs. *as if about to*); (ii) whereas BGT allows for several uses, BAT basically only expresses immediate future (Collins 2009; Höche 2010), one of the earliest meanings of BGT (Eckardt 2006; Garrett 2012), which implies a younger stage of grammaticalization of some future forms (Bybee et al. 1991: 32).

I searched the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) for the strings "modal + BGT" and "modal + BAT". The modal slot was occupied by *will, may, must, should, can, could, might,* and *would*. The cooccurring predicates in the infinitive were classified into activity verbs (including semelfactives), state verbs, and telic verbs (including accomplishments and achievements). Due to the word limitation, I only provide a general picture of the results. 32 occurrences of "modal + BGT" were retrieved from BNC and 70 from COCA, while 65 occurrences of "modal + BAT" were retrieved from BNC and 333 from COCA. The BGT type cooccurred with 47 activity verbs, 39 telic verbs and 11 state verbs, whereas the BAT type appeared with 77 activity verbs, 303 telic verbs and 23 state verbs. In addition, BGT occurred most with *might* (33) and second most with *will* (18); BAT occurred most with *may* (242) and second most with *might* (98).

These observations can be explained along the lines of Wada's (2000, 2019) account of BGT and BAT. BAT is basically restricted to immediate future, whose temporal structure includes the focused part of the situation in the present area, so it strongly tends to occur with present-oriented modals like *may* and *might*. It is not grammaticalized to the stage of modal-like forms (expressing prediction), so it can occur with modals more easily, as quasi-auxiliaries like *have to* can. By contrast, BGT has more uses including future-oriented ones (Leech et al. 2009), which have the respective temporal structures, and thus can go with both future-oriented modals like *will* and present-oriented modals like *might*. It is less likely to cooccur with modals because it is already on the way to the stage of modals (Collins 2009).

References

Brisard, Frank (2001) Be going to: An exercise in grounding. *Journal of Linguistics* 37, 251-285.

Bybee, Joan, William Pagliuca and Revere Perkins (1991) Back to the future. In Elizabeth C. Traugott and Bernd Heine (eds.), *Approaches to Grammaticalization Volume II*, 17-58. Amsterdam/Philadelphia: John Benjamins.

Eckardt, Regine (2006) *Meaning Change in Grammaticalization: An Enquiry into Semantic Reanalysis*. Oxford: Oxford University Press.

Gesuato, Sara and Roberta Facchinetti (2011) GOING TO V vs GOING TO V-ing: Two equivalent patterns? *ICAME Journal* 35, 59-94.

Höche, Silke (2010) What about be about? Walking the tightrope between tense and aspect. *Rice Working Papers in Linguistics* 2, 52-74.

Langacker, Ronald (1990) Subjectification. *Cognitive Linguistics* 1, 5-38.

Mee, Joshua (2013) *The Evolution of Constructions: The Case of* Be About To. MA Thesis at The University of New Mexico.

Nicolle, Steve (1997) A relevance theoretic account of be going to. *Journal of Linguistics* 33, 355-377.

Tagliamonte, Sali A., Mercedes Durhum and Jennifer Smith (2014) Grammaticalization at an early stage: Future be going to in conservative British dialects. *English Language and Linguistics* 18, 75-108.

Wada, Naoaki (2000) Be going to and be about to: Just because Doc Brown was going to take us back to the future does not mean that he was about to do so. *English Linguistics* 17, 386-416.

Wada, Naoaki (2019) *The Grammar of Future Expressions in English*. Tokyo: Kaitakusha.

Watanabe, Takuto (2011) On the development of the immediate future use of be about to in the history of English with special reference to late Modern English. *English Linguistics* 28, 56-90.

Prepositions at the semantics/pragmatics interface: A corpus-based analysis in a cognitive linguistics framework

Michelle Weckermann (University of Augsburg)

Prepositions have been studied extensively in relation to their polysemy, especially through a cognitive linguistic lens. Most of this research has focused on the preposition *over*. The two arguably most influential approaches in this line of research are Lakoff's (1987) full specification account and Tyler and Evans' (2003) principled polysemy account.

As the name suggests, the aim of Lakoff's account is to represent all senses associated with *over*, so that every slight change in meaning receives its own mental representation. This fine-grained approach was heavily criticised for the lack of methodology in defining senses, the sole reliance on introspective judgments, and the use of fabricated examples (see Sandra & Rice, 1995; Tyler & Evans, 2003). Tyler and Evans' (2003) approach drastically reduces the number of senses for *over* in an attempt to define a sense and separate a semantic, conventionalised sense from contextual influences, hence increasing the objectivity of their research. Nevertheless, their approach is still based on fabricated examples and introspection. The present study aims to improve on this research by investigating a range of prepositions (including *over*) and doing so with natural data, methodological criteria for defining a sense (adopted and refined from Tyler & Evans, 2003), and by ensuring inter-coder reliability. Natural data was gathered from a range of data sources, including a legal corpus (*EuroParl*) and four novels from different genres (thriller, romance/drama, dystopia/fantasy, and philosophical novels). The data is thus representative of a range of topic areas, which should mirror as many prepositional senses as possible. The data was coded with respect to the different senses of each preposition and the aforementioned methodological criteria were applied. This was to ensure that a sense is sufficiently different from other, already existing senses (e.g., a different image-schematic configuration or an abstract/metaphorical extension), and that the meaning a sense expresses is not contingent on the sentential context or knowledge supplied by inference (Tyler & Evans, 2003).

Regarding the polysemy of *over*, thirteen senses were identified, including senses of a spatial, temporal, and abstract nature (see (1)-(3) for examples), and then analysed qualitatively and quantitatively. Results show that *over* is primarily spatial (the two spatial senses having the highest frequencies), followed by a sense pertaining to a temporal duration; abstract senses, on the other hand, are much less frequent. Borderline cases (i.e., non-prepositional uses such as *over* as a prefix or in compounds; e.g., *overkill*, *overground*) were also examined in relation to the senses. The senses were further arranged in a semantic network with no central sense but instead clusters of related senses and connections showing the nature of their interrelations (following Rice, 1992, 1993).

(1) I put my wet socks over a radiator to dry. (spatial 'above' sense; Moyes, 95)
(2) She had gone through hundreds over the years (temporal 'duration' sense; Cole, 24)
(3) (…) and the conversation was over. (abstract 'completion' sense; Cole, 28)

References

Cole, D. (2017). *Ragdoll*. London: Orion Books.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.

Moyes, J. (2012). *Me Before You*. Penguin Books.

Rice, S. A. (1993). Far afield in lexical fields: the English prepositions. In Bernstein, M. (ed.) *ESCOL* '92. pp. 206–217.

Rice, S. A. (1992). Polysemy and lexical representation: The case of three English prepositions. *Proceedings of the Fourteenth Annual Conference of The Cognitive Science Society*, pp. 89-94. Routledge.

Sandra, D. & Rice, S. (1995). Network analyses of prepositional meaning: Mirroring whose mind – the linguist's or the language user's? *Cognitive Linguistics* 6(1), pp. 89-130.
Tyler, A. & Evans, V. (2003). *The Semantics of English Prepositions*. Cambridge: Cambridge University Press.

The diachronic change of the *to*-infinitive and gerund as subjects

Michiko Yaguchi (Kanazawa University)

This study aims to quantitatively explore the historical transition in sentence subjects using the *to*-infinitive (e.g., "*To see is to believe*") and gerund (e.g., "*Seeing is believing*") forms. While previous studies have discussed the general trends over time for these subject types (e.g. Visser 1966: 942-971, 1098-1102), they have not clarified detailed changes regarding their frequencies or how their functional characteristics have evolved. This study addresses these gaps, examining British English data from Early Modern English to contemporary British English in the 2000s. The analysis is conducted using the Helsinki Corpus, Corpus of Late Modern English Texts, LOB corpus, FLOB corpus, and British English 2006 (BE06). Additionally, the study extends to present-day American English contexts, examining the Brown corpus, Frown corpus, and American English 2006. The analyses unveil two findings regarding the diachronic changes in these two subject forms.

Firstly, the changing frequencies of the two subject types over time are presented. In the Early Modern English period, they showed similar frequencies. However, subjects using the *to*-infinitive reached their highest frequency in Late Modern English, while those using the gerund form were rarely employed during the same period, largely due to prescriptive grammar. In contemporary English, the trend has reversed: the use of the former has considerably decreased in frequency, in sharp contrast to the significant rise in the latter. This shift has led to a ratio of 1:10 between the former and the latter in current English, as observed in the 2000s in both British English and American English.

Secondly, there is a notable difference in the types of constructions between these two forms synchronically as well as diachronically. Concerning the *to*-infinitive as subject, three specific constructions have consistently represented over half of all its instances since the 1990s. These three constructions are: (1) the *to-to* construction (e.g., "*...but just to pose them is also to raise the suspicion...*" BE06), (2) the *to-would* construction (e.g., "*To take away jobs in a town like Launceston would be a huge blow.*" BE06), and (3) the *to-require* construction (e.g., "*To answer this question requires commentary on MacIntyre's notion of virtue*..." BE06). Indeed, the proportion of these three constructions to the total occurrences of the *to*-infinitive as subject has been constantly high (30% or more) since Early Modern English, and this trend has persisted until now. This suggests that, rather than expanding in productivity, the *to*-infinitive form has been moving toward the limited use in these three constructions, along with the significant decrease in frequency since the 1990s especially.

In contrast, the gerund form has commonly been observed in various constructions through Modern English to present-day English, including sentences expressing achievement (e.g., "*Writing this book has given me the opportunity to stand back...*", BE06). This kind of construction is rarely seen with the *to*-infinitive form in present-day English.

References

Duffley, Patrick J. (2003). The gerund and the *to*-infinitive as subjects. *Journal of English Linguistics* 31: 324-52.

Rohdenburg, Günter. (2006). The role of functional constraints in the evolution of the English complementation system. In *Syntax, Style and Grammatical Norms. English from 1500-2000*, ed. by C. Dalton-Puffer, D. Kastovsky and N. Ritt, pp. 143-166. Bern: Peter Lang.
Swan, Michael. (2017). *Practical English Usage*, fourth edition. Oxford: Oxford University Press.
Visser, Fredericus T. (1966). *An Historical Syntax of the English Language,* Part II. E. London: J. Brill.
Yaguchi, Michiko. (2023). Development of the subject of BE *going to* in grammaticalisation from the 1820s to 2010s in comparison with BE *about to*. *English Studies* 104(7): 203-224.

Prodigy team as a game changer in computational corpus annotation

Wenwen Guan (University of Amsterdam)

Corpus annotation is an essential pre-requisite for corpus-based linguistic research and natural language processing (NLP) tasks. The latest advances in AI and NLP have also contributed to the development of corpus technology. For instance, automated annotation of form-based annotation tasks including phonetic annotation, part-of-speech (POS) tagging, and syntactic parsing. By contrast, automated annotation has encountered a bottleneck in its application to functional linguistic terms, particularly in the fields of pragmatic and discourse analysis (Hovy & Lavid, 2010; Lu, 2024). Linguists would rather endure the intensive labour that those annotations demand when anticipating the difficulty in programming or developing a specific tool. Inspiringly, an all-in-one annotation tool called Prodigy Teams (PT) promises significant progress in automation of function-based annotation and enables linguists who have little programming knowledge to have fun with NLP resources.

A metadiscourse annotation task will be used as an example to demonstrate how PT works. The task is tricky due to the context-dependence and multifunctionality of metadiscourse. These features complicate the formulation of annotation guidelines and training of human annotators. In PT, the complicatedness is properly handled by a highly interactive user interface (UI), where you can manage the annotator team, provide guidelines, monitor annotation progress, review the labels and so on. The UI can be easily invoked with PT's hands-on instructions, and it makes metadiscourse annotation and review as handy as swiping spans in the texts. More importantly, it is the prediction function that makes PT stand out from numerous corpus tools. On the one hand, PT has a rule-based matching function which takes advantage of existing annotated resources such as POS tags and regular expressions to identify form-based types of metadiscourse. On the one hand, PT features in an iterative machine learning (ML) loop. The loop can start with a tiny manually annotated dataset, for example, only 10,000 tokens, and predict labels for the rest of the raw data. When training the ML model, PT can evaluate the model performance and display the prediction accuracy. After another 10,000 tokens, the model can be fine-tuned to predict more accurate labels. The process can be repeated multiple times until producing a satisfactory accuracy. In a primary experiment, the model achieved an accuracy of 0.77 after being trained with only 89,000 tokens. It indicates that the loop can undoubtedly speed up annotating efficiency and produce more accurate labels.

To sum up, PT significantly simplifies the annotation process and ensures annotation consistency by incorporating all necessary functions on one platform. In addition to metadiscourse annotation, PT can also be customized for various annotation purposes, for instance, dependency & relations annotation, multimodal data annotation, LLM applications, and even their combinations. The demonstration caters to the interests of corpus builders, especially who are working on understudied linguistic objects and self-defined annotation schemes with limited coding skills. The workflow and the source codes for post-annotation data processing will be published and shared through my github.

References

Hovy, E. & Lavid, J. (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation* 22, 13-36.

Lu, X. (2014). *Computational Methods for Corpus Annotation and Analysis*. Springer Netherlands. https://doi.org/10.1007/978-94-017-8645-4

A voice-based Chatbot for Language Learners (ChaLL)

Gerold Schneider[1], Johannes Graën[1], Manuela Hürlimann[2], Luzia Sauer[3], Janick Michot[2], Mark Cieliebak[2] and Matthias Bachmann[2]

([1]University of Zurich, [2]Zurich University of Applied Science, [3]Zurich University of Teacher Education)

We present a prototype of ChaLL, a voice-based chatbot that provides language learners with opportunities to practice speaking English in both focused and unfocused task-based conversations and receive feedback, free from the time constraints and pressures of the traditional classroom setting. We address pupils between about 10 and 15 years, in classroom settings or in their spare time.

Speaking practice is essential for successful foreign language learning; however, it can be difficult to achieve this in the classroom: often, there is not enough time to allow for all learners to speak, and fear of being judged can make it difficult to speak freely. We have developed a prototype of ChaLL (Chatbot for Language Learners), a voice-based chatbot that will provide learners with vital speaking opportunities.

In order to become the ideal companion for improving learners' fluency, ChaLL needs to

- Adjust its level of speech complexity, including grammar and vocabulary, to learner levels to ensure that the interactions are in the "zone of proximal development", which is the optimal level for of potential development.
- Provide adequate real-time feedback, e.g. by including a proper version of the non-standard input into the spoken response (that is recasting), and support to achieve an ideal learning effect.
- Give useful feedback, be entertaining and encouraging to increase learners' self-confidence.

We first introduce the transformer-based methods that ChaLL depends on:

1) Speech-To-Text technology to adequately "understand" learners' free speech – including the errors they make and non-standard language use – as the basis for interacting with the chatbot. Transformer-based systems like OpenAI Whisper have revolutionized text-to-speech methods (Radford et al. 2022). We give an evaluation in terms of word error rates.
2) Automatically detecting and classifying errors in the automatically transcribed speech as the basis for providing feedback. We rely on Bryant et al. (2019) and Schneider (2023).
3) The degree to which learners' skill levels in different dimensions (e.g. grammar, lexical choice or pronounciation) can be identified automatically.
4) State-of-the-art transformers, in particular ChatGPT (OpenAI 2023), which helps us to create a chatbot that can have meaningful and entertaining conversations with pupils.

Second, we give a live demonstration of ChaLL, in order to illustrate the power and the current weaknesses of our system. The system is entertaining and recognizes errors fairly well, but latency, mispronunciation and the use of non-English words (in the pupils' native language) still pose challenges.

References

Bryant, Christopher, Felice, Mariano, Andersen, Øistein E., and Briscoe, Ted. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 52-75. https://aclanthology.org/W19-4406

OpenAI. 2023. *ChatGPT* (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavey, Christine, and Sutskever, Ilya. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv*:2212.04356 [eess.AS]. https://arxiv.org/abs/2212.04356

Schneider, Gerold. 2023. Detecting and Analysing Learner Difficulties using a Learner Corpus Without Error Tagging. In Kieran Harrington and Patricia Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching* (pp. 229-257). London: Palgrave Macmillan/Springer.

Acknowledgements

We gratefully acknowledge the financial support of the following institutions/funding bodies: Xunta de Galicia (Regional Government); Language Variation and Textual Categorisation (LVTC) research group; Consello Social (Universidade de Vigo); Department of Filoloxía Inglesa, Francesa e Alemá (FIFA); Concello de Vigo and Turismo de Vigo (Vigo City Council, Council's Tourism Department).

We are also grateful to the publishing houses which generously contribute to this conference in various ways (alphabetical order): Brill, Cambridge University Press, John Benjamins, Lincom, Multilingual Matters.