Abstracts ICAME44

Vanderbijlpark, 17-21 May 2023

Sequence: keynote speakers, then workshop speakers (17 May), then parallel session papers (18-21 May), all in alphabetical order, although probably most easily searched electronically.

Keynote Speakers

Challenges of doing corpus linguistics in Africa

Alexandra Esimaje

Language research has become of pre-eminent importance now than it was decades ago when the relevance of language to society was less discernible. There are many societal problems whose origin bother on language whether directly or indirectly. The necessity of resolving such problems has put substantial burden on the conventional and introspective methods of language research. Corpus linguistics seems to be solving a great part of this problem by presenting itself as an alternative method of linguistic investigation; offering new perspectives on language related issues, new ways of engagements and producing significant results. Since its inception in the early 1990s, its effect has been nothing but phenomenal because it has enabled investigations at depths and speeds only imagined before now and its outcomes have proved themselves outstanding. This is why the corpus method has continued to hold sway in language academies in most parts of the global North where it originated. In those places, there are many established corpora which are relatively easily accessible, many software exist that are fairly commonly known and used by linguists across universities, and a great number of published outcomes of corpus research are found to be in application in many contexts, whether directly in the classrooms or applied in the real world.

This is not the case in the global South, especially in many parts of Africa, where corpus linguistics remains largely unknown and the knowledge it provides even more largely untapped. Given that online technologies such as the Internet have made considerable knowledge only a finger click away, the presumed unfamiliarity with corpus linguistics manifested by many language teachers/researchers in the South should be a valid concern for the global linguistic community. The state of corpus linguistics in Africa is therefore the prime interest of this address. It seeks to define and illustrate the extent of its knowledge, use, outcomes, and impact, and also to examine the challenges facing its growth as a field of study and adoption as a reliable method of linguistic enquiry. Through ethnographic surveys of corpus linguistics studies across the countries of: Nigeria and Ghana in West Africa; Tanzania, Uganda, and Kenya in East Africa; Zimbabwe and South Africa in Southern Africa; and Cameroon in Central Africa, this paper provides some insights into the state of corpus research in Africa and reveals the challenges facing it. The conclusion that forces its way through is that in Africa, corpus linguistics may remain ‘relatively new’, not in terms of age nor in terms of its appropriateness but rather in terms of its feasibility or usability, due to the many limiting factors of underdevelopment and the management of higher education in the continent. This calls for collaborative interventions if the field is to be sustained and its impact globally recognized.

Crossing Spaces in Learner Corpus Research

Sylviane Granger

The field of learner corpus research (LCR), which emerged in the late 1980s/early 1990s, aimed to add to corpus linguistics one variety that was not yet covered at the time, that of learner English, i.e. written and/or spoken data produced by foreign or second (L2) language learners. The objectives of this type of research were – and still are – to gain a better understanding of the process of learning a foreign or second language and to design more efficient language teaching and testing tools and methods. Learner corpus research is now an established subfield within both corpus linguistics and second language acquisition, with its own scientific association, journal, biennial conference series and dedicated handbook (Granger et al. 2015).

In its over thirty years of existence the field has remained true to its initial objectives, but it has progressively outgrown its initial boundaries and is still expanding. In my presentation, in keeping with the theme of the conference "English going places, Corpora crossing spaces", I will tackle in turn three types of boundaries that LCR has crossed over the years: language boundaries, disciplinary boundaries and conceptual boundaries.

In terms of language, the field has expanded from a near-exclusive focus on L2 English to an impressive number of other L2s. This expansion has been accompanied by an equally large increase in the number of learners' mother tongue backgrounds covered. One weakness of the majority of current studies is that they fail to take into account the acquisition setting. Apart from some rare exceptions, no distinction is made between English as a Foreign Language learners and English as a Second Language learners, thereby neglecting the degree of exposure to the target language.

The number of disciplines to which learner corpus research is linked has also expanded considerably. The initial synergies with Second Language Acquisition and Foreign Language Teaching have been progressively complemented with links to other crosslingual varieties (Granger 2015), i.e. language varieties that have specific characteristics due to the interplay of two or more languages, in particular Contrastive Linguistics and Translation Studies, World Englishes (WE) and English as a Lingua Franca (ELF).

Conceptually, the notions of 'native speaker' and 'error', which have been at the heart of LCR since its origin, have been revisited following criticisms from several quarters, in particular WEs (Tan 2005) and ELF (Seidlhofer 2001). I will argue, however, that used advisedly, in full awareness of their limitations and in keeping with the research objective pursued, these notions still have a key role to play in learner corpus research and should therefore not be overly stigmatized.

Reference

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1), 7-24.

Granger, S., Gilquin, G. & Meunier, F. (eds) (2015). *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press.

Seidlhofer, B. (2001). Closing a conceptual gap: the case for a description of English as a lingua franca. *International Journal of Applied Linguistics* 11(2), 133-158.

Tan, M. (2005). Authentic language or language errors? Lessons from a learner corpus. *ELT Journal* 59(2), 126–134.

Does editing matter? Editorial practice as a factor in language variation and change in written L1 and L2 varieties of English

Haidee Kotze

The potential influence of editorial intervention on published texts produced by users of L1 and L2 varieties of English is often commented on in passing in the context of reflections on the degree to which corpora reflect ‘authentic’ usage. At the same time, the degree to which innovative forms are regarded as acceptable by gatekeepers of the publishing industry, like editors, is frequently raised as a measure of endonormativity in L2 varieties of English. However, substantive empirical and specifically corpus-based research investigating editorial intervention and its implications, across varieties of English, is limited. In this presentation, I outline a rationale for why editing matters in studies of language variation and change, and present a usage-based model for the role of editing as a factor that both *reflects* and *affects* endonormativity and convergence in varieties of English. This is followed by an overview of some recent corpus-based work in this area, with a focus on South African Englishes.

Corpora crossing disciplinary borders - ICE Uganda meets macro-sociolinguistics, anthropology and ethnography

Christiane Meierkord

Over the last decades, there has been an enormous increase of corpus linguistic studies in the field of world Englishes, following the availability of a growing number of components of the *International Corpus of English* (ICE; Greenbaum 1996, Kirk & Nelson 2018). They have provided the research community with important cross-variety analyses of various grammatical and lexical characteristics as well as of phonological and pragmatic features, often paired with increasingly sophisticated inferential statistics and their visualisation.

However, the focus on significance testing (potentially due to the security associated with *p* values) may have come at the expense of neglecting to integrate aspects that further our understanding of *why* Englishes around the world differ in the ways they do and that trigger related research questions. In fact, world Englishes research has traditionally integrated social and cultural factors in descriptions of the English language complex (see e.g. the papers in Kachru 1982).

This paper presents three studies conducted with the ICE-Uganda (Meierkord & Isingoma 2022) data and describes how their results interpretation benefited from appreciating the individual ICE texts from within their immediate, social and cultural contexts. It starts from a discussion of currently existing research that crosses theoretical as well as methodological borders between corpus linguistics on the one side and macro-sociolinguistics, anthropological linguistics and linguistic ethnography, before introducing Uganda and ICE-Uganda.

Thereafter, results from investigations into the assumed Kiswahili influence on the Ugandan English lexicon, into praise as a speech act prevalent in Uganda, and into expressions of future time across Ugandan English speakers with different language family backgrounds will be presented. It was found that the low social status of Kiswahili renders the Ugandan English lexicon considerably different from that of Kenyan and Tanzanian English. Persisting strong interdependencies across individuals in Uganda demand that these be acknowledged, for example through praise, and the sociolinguistic structure found in the country's many boarding schools results in accommodation to the variety of English spoken by speakers with a Bantu language background.

References

Greenbaum, Sidney (ed) (1996). *Comparing English Worldwide. The International Corpus of English.* Oxford: Clarendon Press

Kachru, Braj B. (ed.) (1982). *The Other Tongue: English across Cultures*. Urbana-Champaign, IL: University of Illinois Press.

Kirk, John & Gerald Nelson (2018). The International Corpus of English project: A progress report. *World Englishes*, 37(4). 697-716.

Meierkord, Christiane & Bebwa Isingoma (2022). The International Corpus of English - Uganda (ICE-UG). www.rub.de/engling/research/uganda/corpus.html.en

News Downloads and Text Coverage: Case Studies in Relevance

Mike Scott

With the increasing availability of huge text databases, it has become easier for researchers in institutions with library subscriptions to access large amounts of text to process in studying a research topic. This has radically changed the data usage pattern for not only the academic world but also those engaged in politics, medicine, history, journalism. Such a corpus is not designed to represent the characteristic features of a language but the characteristic treatment of a topic by the data sources in question. Corpora crossing spaces, many spaces indeed!

However the issue of just how relevant the texts which can be downloaded are to the search-terms used in finding them has not received enough attention. With tens or hundreds of thousands of texts in a study there's no question of reading each one to check its relevance. For some studies it may not be appropriate to worry how relevant the search-terms are to the text because incidental and very localised usages may still be of interest. Some topics may, one imagines, be of their nature incidental and not something many texts are genuinely about. For other research aims, though, it may be worth attempting to filter a complete download in an attempt to find only those texts which are truly about the topic in question. Grundmann & Scott 2014, studying climate change, for example, wished to find out what is being said and by whom about global heating, carbon footprint etc., ignoring texts where there was only a passing reference to climate or a joke about the weather. In the case of Grundmann et al 2017, the aim was to filter austerity texts so as to exclude not only passing references but also austerity in Greece which was prominent in the downloads of the time. Working with such corpora shocked me to see how many irrelevant texts there were in the corpus and led to the current study.

This presentation deals with on-going research into the issue. Research questions included a) What is meant by *relevance* in news text downloads on *water, climate, austerity* etc?, b) What can a key words database show us about the nature of such topics?, c) How highly do texts score on relevance?, d) How reliable are automated relevance scores? Materials are downloads of over 130,000 news texts on a variety of topics. The presentation will show the procedures and the current analysis of the breakdown of relevance scores.

References

Grundmann, Reiner & M. Scott 2014. Disputed climate science in the media: Do countries matter? *Public Understanding of Science*. DOI: Vol. 23(2) 220 –235.

Grundmann, Reiner; Kreischer, Kim-Sue and Mike Scott. 2017. The Discourse of Austerity in the British Press. In: Roland Sturm, Tim Griebel, Thorsten Winkelmann (eds.), *Zeitschrift für Politik Special Issue 8: Austerity: A Journey to an Unknown Territory*, 92 – 128.

Pre-conference workshop 17 May 2023

Corpora crossing language borders

Going places: English GO and Norwegian GÅ – mutual correspondence and textual variation

Signe Oksefjell Ebeling and Hilde Hasselgård

Taking two cognates in English and Norwegian as its point of departure, this paper has two main aims: (i) to expand and nuance the procedures of calculating translation bias (TB) and Mutual Correspondence (MC) (Altenberg 1999) in contrastive studies; and (ii) to carry out a qualitative contrastive analysis of GO and GÅ.

GO and GÅ are syntactically and semantically highly versatile, and have been shown to have both overlapping and non-overlapping uses in English and Norwegian/Swedish (see e.g. Viberg 1999, Cej 2008). Due to their multifunctional nature, they are very frequent verbs in both languages, and thus suitable for an experimental study that relies on widely used and widely dispersed items.

The TB and MC of items in two languages are calculated on the basis of bidirectional corpus data, where the number of times the items are translated into one another in a given corpus results in a percentage of correspondences: unidirectional (TB) or bidirectional (MC). Both measures are intended to show how similar, or mutually translatable, two linguistic items are. Traditionally, this procedure has been carried out without attending to variation, i.e. dispersion, across the different corpus texts. However, we believe that the measures will become more reliable if textual variation is taken into account, while the conclusions based on them will be more robust.

As a case study, we analyse the lexemes GO and GÅ and their translations in the fiction part of the English-Norwegian Parallel Corpus (30 text extracts). The analysis considers lexicogrammatical features such as verb form and syntactic pattern, as well as congruence in translation, and thus differs from Viberg (1999). The semi-auxiliary BE *going to* is excluded from the analysis (unlike Cej 2008), while idiomatic combinations such as GO *in for* and GÅ *an* are included, but marked as phrasal. The proportion of congruent translations – i.e. instances where GO and GÅ correspond to each other in translation – is calculated per text, and can also be broken down according to verb form or syntactic pattern.
In a first step, we analysed the base forms *go* and *gå* and focused on the infinitive forms of intransitive, non-phrasal uses, as in example (1), where the translation is fully congruent, i.e. *gå* corresponds to *go*, in the same syntactic pattern.

(1) Herman liker å *gå* dit etter skoletid ... (LSC1)
Herman likes to *go* there after school ... (LSC1T)

Fully congruent translations were about equally frequent in both translation directions at group level, but the individual texts/translators varied substantially, from 0% to over 70% congruence. Other patterns are expected to show less congruence, especially phrasal uses, as in (2), but textual variation is expected for such patterns too.

(2) He would *go on* now for weeks ... (MW1)
Han ville *fortsette* nå i ukesvis … (MW1T). (lit.: ‘continue’)

When the complete material has been analysed, we will address the following questions:

i. Can translation correspondence be better illuminated when taking textual variation into account?
ii. Can translation correspondence be better explained if based on lexicogrammatical features, and taking textual variation into account?

References

Altenberg, Bengt. 1999. Adverbial Connectors in English and Swedish: Semantic and lexical correspondences. In *Out of Corpora. Studies in Honour of Stig Johansson,* edited by Hilde Hasselgård and Signe Oksefjell, 249-268. Amsterdam: Rodopi

Cej, Alla. 2008. The polysemous cognates English *go*, German *gehen* and Norwegian *gå*: a corpus-based contrastive study. MA thesis, University of Oslo. http://urn.nb.no/URN:NBN:no-21406

Viberg, Åke. 1999. The polysemous cognates Swedish *gå* and English *go*. Universal and language-specific characteristics. *Languages in Contrast* 2, 89–115

Putting something somewhere in English and Norwegian: a contrastive approach

Thomas Egan

This paper deals with the English verb *put* in *Subject – Verb – Object – Adverbial* (SVOA) constructions (Quirk et al. 1985: 63) and its Norwegian correspondences, based on data from the English source and target texts, both fictional and non-fictional, in the English–Norwegian Parallel Corpus. In the prototypical sense of the SVOA construction in both languages, the Subject (S) encodes an AGENT, who exerts force on a concrete THEME, coded by the Direct Object (O), causing it to move to a GOAL, encoded by the Adverbial (A).

English *put* is a most versatile verb, represented by 56 separate head definitions in the OED. It is the first caused-motion verb used by small children (Goldberg 2005: 109), and the second most common one (after *take*) in conversation among adults (Biber et al. 1999: 367). It occurs with a wide variety of particles (intransitive prepositions), and many of these, such as *put up* and *put on*, occur in caused-motion constructions.

In the SVOA construction *put* does not itself denote either the PATH along which the THEME is moved, or the GOAL. Nor does it denote the MANNER of the force exerted by the AGENT. Predications containing *put* may, however, specify either (part of) the PATH (*put it down*), the GOAL (*put it on the table*) , or both PATH and GOAL (*put it down on the table*). The two Norwegian verbs *legge* and *sette*, which together account for 53% of the Norwegian translations in the ENPC of SVOA *put* predications with concrete affected THEMEs, resemble *put* in encoding either the PATH, the GOAL, or both.

According to Viberg (2015) the three postural placement verbs *sätte*, *ställe* and *lägge* account for 68% of Swedish translations of caused-motion *put* predications in his corpus. He mentions the Norwegian posture placement verb *stille* as corresponding to Swedish *ställe*, but states that '*sette* seems to have a strong tendency to replace *stille*', just as *zetten* has replaced *stellen* in Dutch (Lemmens 2002: 262). That this tendency is very strong indeed is shown by the fact that, of the 285 relevant examples containing *put* in the English source texts in the ENPC, which are translated by 45 different verbs, just one example is translated by *stille.*

In this paper I examine all tokens of *put* in SVOA constructions in the ENPC, and address the following three research questions:

(1) (How) does the semantics of the PATH influence the choice of the caused-motion verb used in translations of *put* into Norwegian?
(2) (How) do the semantics of the THEME and GOAL (whether they denote animates or body parts, for example) influence the choice of verb or preposition used in the translations?
(3) What characterises examples that are translated by verbs other than *legge* and *sette*?

References

Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad and Edward Finegan. 1999. *Grammar of Spoken and Written English*. John Benjamins.

Goldberg, Adele. 2006. *Constructions at Work: the nature of generalization in language*. Oxford University Press.

Lemmens, Maarten. 2006. Caused posture: experiential patterns emerging from corpus research. In Aanatol Stefanowitsch and Stefan Gries (eds), *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis,* 261-296. Mouton de Gruyter.

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

Viberg, Åke. 2015. Contrasts in construction and semantic composition: The verbs of putting in English and Swedish in an intra-typological perspective. In Signe Oksefjell Ebeling and Hilde Hasselgård (eds), *Cross-Linguistic Perspectives on Verb Constructions*, 222-253. Cambridge Scholars Publishing.

Recapitulating discourse markers in English, Spanish and Spanish translated from English

Camino Gutiérrez-Lanza and Rosa Rabadán

Bilingual and multilingual corpora have been and are essential in understanding the relationships between languages. They help to grasp the internal workings of each of the languages represented. Additionally, they allow us to investigate whether language boundaries are affected by cross-linguistic mediation (Čermáková et al. 2021). One area where English and Spanish tend to behave differently is the marking of discourse relations (Aijmer et al. 2006, Rabadán and Gutiérrez-Lanza 2023 in press). This presentation focuses on one type of non-paraphrastic reformulating discourse markers (DMs): recapitulating DMs (Martín Zorraquino and Portolés Lázaro 1999, Cuenca 2003, Cuenca and Back 2007, Del Saz 2007, Garcés Gómez 2017, Ruiz González 2020, Murillo Ornat 2021). These DMs signal that the text following is a conclusion or a summary of a previous argument and may engage in the same or a different argumentative line (Garcés Gómez 2003 and 2005, Borreguero Zuloaga 2015).

This paper intends to unveil the similarities and differences a) between English and Spanish when recapitulating, both in fiction (F) and in non-fiction (NF), and b) between translated and non-translated DM Spanish usage. English DMs *in summary/in sum/in conclusion; in general, generally; after all,* have been found to trigger formal interference in Spanish and analysed separately (Rabadán and Gutiérrez-Lanza 2023 in press). To avoid the effects of word-for-word translation, here we have chosen recapitulating DMs unaffected by formal interference: *en fin* (in the end)*, en resumidas cuentas* (in a nutshell), *a fin de cuentas (*after all*), en definitiva* (ultimately*), definitivamente* (definitely)*,* and *total* (I mean) *(*Martín Zorraquino & Portolés Lázaro 1999: 4051-4143).

To carry out the analysis, we use three corpora: a bilingual parallel corpus (English-Spanish P-ACTRES 2.0), a monolingual corpus of Spanish translated from English (CETRI), and a monolingual reference corpus of original Spanish (CORPES XXI). Concerning corpus comparability, the three corpora include the same type of fiction and non-fiction materials and, in the case of Spanish, feature the same geographical variety, i.e., European Spanish. Since they are different in size, inferential statistics have been applied to ensure the statistical significance of the results. We searched our selected DMs in the English-Spanish parallel corpus (P-ACTRES 2.0) to identify English triggers and their frequencies. Then, original English and Spanish data are contrasted to determine whether both languages recapitulate similarly or otherwise. Next, translated DMs are queried in CETRI (translated Spanish) and CORPES XXI (non-translated Spanish) and the results are compared to showcase (non)significant differences.

English language results show that the chosen recapitulating DMs most frequently derive from *well* (27%), *anyway* (16.2%), *you know/you see* (9.5%) *in short* (5.4%), *I mean* (2.7%) and *after all* (2.7%). Additionally, 27 % of the DMs in the translations do not have a trigger in English. Non-translated results indicate that our chosen DMs signal recapitulation more frequently in Spanish than in English. The frequencies are in F, 38.38 per million words (pmw) in English vs 93.96 pmw in Spanish; in NF: 6.34 pmw in English vs 81.22 pmw in Spanish. Translated and non-translated Spanish results suggest that translated usage of recapitulating DM *en fin* (in the end) is normalized, as it is the most frequent DM in translation and non-translation. Different trends can also be observed concerning distribution: a) Deflation, i.e., the most frequent DMs in non-translated Spanish are underused in translation both in F (*en fin, en definitiva, total, a fin de cuentas*) and in NF (*en definitiva, en fin*), thus reflecting English usage; b) Dilation, i.e., the less frequent DMs in non-translated Spanish (*definitivamente, en* *resumidas cuentas*) are overused in the NF translations, but they show no significant difference in the F translations (equalization). The results highlight the role of corpora in tracking trends across language boundaries. Further work includes comparing these findings to those of non-interfered recapitulating DMs.

References

Aijmer, K. et al. 2006. Pragmatic markers in translation: A methodological proposal. In Fischer, K. (ed.). *Approaches to Discourse Particle*s. Amsterdam: Elsevier. 101-114.

Borreguero Zuloaga, M. 2015. A vueltas con los marcadores del discurso: de nuevo sobre su definición y sus funciones. In Ferrari, A. & Lala, L. (eds). *Testualità. Fondamenti, unità, relazioni*. Firenze: Franco Cesati. 151-170.

Čermáková, A., Ebeling, S.O., Levin, M. & Ström Herold, J. (eds) (2021). Crossing The Borders: Analysing Complex Contrastive Data. *Bergen Language and Linguistics Studies (BeLLS*), 11(1). https://doi.org/10.15845/bells.v11i1.

Cuenca, M. J. 2003. Two ways to reformulate. A contrastive analysis of reformulation markers. *Journal of Pragmatics* 35(7). 1069-1093.
https://doi.org/10.1016/S0378-2166(03)00004-3

Cuenca, M. J. & Bach, C. 2007. Contrasting the form and use of reformulation markers. *Discourse Studies* 9(2). 149-175. https://doi.org/10.1177/1461445607075347

Del Saz, M., M. 2007. *English Discourse Markers of Reformulation*. Lausanne, Switzerland: Peter Lang.

Garcés Gómez, M. del P. 2003. Los marcadores de recapitulación y de reconsideración en el discurso. *Revista de investigación lingüística* 1(VI). 111-141. https://revistas.um.es/ril/article/view/5531/5391

Garcés Gómez, M. del P. 2005. Reformulación y marcadores de reformulación. In Casado Velarde, M., González Ruiz, R., Loureda Lamas, Ó. (eds.). *Estudios sobre lo metalingüístico (en español)*. Frankfurt: Peter Lang. 47-66.

Garcés Gómez, M. del P. 2017. La reformulación discursiva y los procesos de recapitulación y conclusión. A propósito de los marcadores *en fin* y *total*. *Romanische Forschungen* 129(3). 295-316. DOI: https://doi.org/10.3196/003581217821694319

Martín Zorraquino, M. A. & Portolés Lázaro, J. 1999. Los marcadores del discurso. In Bosque, I. & Demonte, V. (eds.), *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa Calpe. 4051-4213.

Murillo Ornat, S. 2021. Reformulation Markers in Non-initial Position in Written English and Spanish. *Complutense Journal of English Studies* 29. 35-48. DOI: https://doi.org/10.5209/cjes.77793

Rabadán, R. and Gutiérrez-Lanza, C. 2023 in press. Interference, explicitation, implicitation and normalization in third code Spanish: Evidence from discourse markers. *Across Languages and Cultures.*

Ruiz González, N. 2020. Los reformuladores de recapitulación en el corpus Preseea de Granada. *ELUA* 34. 193-212. DOI: https://doi.org/10.14198/ELUA2020.34.9

English and Czech venitive verbs in contrast: deictic, or not?

Michaela Martinková and Markéta Janebová

Languages differ as to which type of deictic centre venitive verbs accept: while *come* codes motion towards the speaker or hearer at coding (utterance) or reference (event) time, their homebase, or the location of the central character in the narrative (Fillmore 1997), Spanish *venir* can only be used for motion towards the speaker. In Slavic languages, the situation is less clear; if Slavic verbs code deixis, then on prefixes. According to Filipović (2010, 253), "[t]he majority of OD-/DO-verbs ['from'-/'to'-verbs] are used deictically," Slobin (2004, 8) considers the Russian *pri-* to be "a deictic prefix on a motion verb," and according to Malá (2015, 174), its Czech cognate *při-* indicates "directed motion towards the deictic centre (the speaker)." Lewandowski (2014, 44), however, argues that distribution of (most of) Slavic equivalents of *come* and *go* "is related to other, non-deictic factors," e.g. Slavic *come* verbs are "preferred when the speaker wishes to adopt an arrival-oriented perspective" (see also Janda et al. [2013] about the Russian *pri-*).

This study adopts a contrastive approach to find out whether Czech *come* verbs are indeed deictic. First, we identify the degree of functional equivalence between *come* and the Czech intransitive motion verbs prefixed by *při-* by calculating Mutual correspondence (MC) between the items in a bidirectional parallel corpus of subtitles (created within InterCorp [Čermák and Rosen 2012]), and then investigate their correspondences. Preliminary results suggest that MC is only 27.9%, with a translation bias: 45.1% of *při*-verbs are translated by *come*, but only 20% of *come* are translated by *při*-verbs. While *při*-verbs are typically used to describe motion to the speaker or hearer, in many such situations they cannot be used, e.g. in those incompatible with the ARRIVAL perspective: (1) calls for the imperative *pojď* [DEIX-walk]; arguably, *po-* in imperatives of motion verbs modifies the meaning of the verb so that it invites the hearer to accompany the speaker, or to move in his/her direction (Petr et al. 1986, 418). *Pojď* or unprefixed imperfectives are the only option in comitative contexts, and the latter are also preferred in situations involving motion-in-progress (2).

(1) *Come to momma!* (IC:SUBTITLES_4027439)
Pojď k mamince!
DEIX-walk-IMP toward mummy

(2) *Coming, sweetheart.* [IC:SUBTITLES_3625781]
Už jdu zlato.
already walk-IPFV.1SG sweetheart

In the opposite direction of translation, *come* is the most frequent translation of *při*-verbs and *go* is rare (1.4%), but is attested; furthermore, (3) has no deictic centre:

(3) *každý, kdo přijede do Rumunska* [IC:SUBTITLES_3371565]
everybody who *při*-go.by.vehicle-2PL to Romania
(It's clear from the briefing notes that) anyone who goes to Romania (gets the same doctorate as me.)

Even these preliminary results suggest that the presence of a deictic centre is expected but not necessary for the use of *při*-verbs; in future this will need to be tested on a larger set of data as well as experimentally. Since cross-linguistic differences may lead to interference in L2 interlanguage as well as in translation, their identification is necessary for future studies of thinking for speaking (Slobin 1996) in L2 and for translating.

References

Čermák, František, and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3), 411–427.

Filipović, Luna. 2010. The Importance of being a prefix: Prefixal morphology and the lexicalization of motion verbs in Serbo-Croatian. In Hasko, V. and R. Perelmutter (eds), *New Approaches to Slavic Verbs of Motion.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Fillmore, Charles. J. 1997. Coming and going. *Lectures on Deixis.* Stanford: CSLI.4

Janda, Laura A., Endresen, Anna, Kuznetsova, Julia, Lyashevskaya, Olga, Makarova, Anastasia, Nesset, Tore, Sokolova, Svetlana. 2013. *Why Russian aspectual prefixes aren't empty: prefixes as verb classifiers.* Bloomington, IN: Slavica Publishers.

Lewandowski, Wojciech. 2014. Deictic Verbs: Typology, Thinking for Speaking and SLA. *SKY Journal of Linguistics* 27, 43–65.

Malá, Markéta. 2015. Translation counterparts as indicators of the boundaries of units of meaning: A contrastive view of the position of "come V-ing" among the patterns of the verb come. In Ebeling, S. O. and H. Hasselgard (eds), *Cross-Linguistic Perspectives on Verb Constructions*. Cambridge Scholars Publishing.

Petr, Jan, Komárek, Miroslav, Kořenský, Jan, and Jarmila Veselková. 1986. *Mluvnice češtiny* [2]. Praha: Academia.

Slobin, Dan I. 1996. From "Thought and Language" to "Thinking for Speaking". In J. Gumperz and S. Levinson (eds.), *Rethinking Linguistic Relativity*. Cambridge University Press.

Slobin, Dan I. 2004. The many ways to search for a frog: Linguistic typology and the expression of motion events. In S. Strömqvist S. and L. Verhoeven (eds.), *Relating events in narrative: Typological and contextual perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Non-verbal plural number concord – a pilot study comparing English and German

Karolina Rudnicka

Non-verbal plural number agreement is, especially in the cross-linguistic context, an under-researched topic. English and German seem to differ with regard to number preference in objects and adverbials (coreferential terms) following plural subjects (antecedents). While English prefers the distributive plural – the agreement in number between the (formally or notionally) plural subject of a clause and a nominal clause element in the predicate part of this clause (Quirk, 1985; Aarts et al., 2014), German seems to be much more open to variation and may even appear to prefer the distributive singular (cf. Duden[1]). Let us compare the examples (1) and (2), which illustrate the typical number of the noun *life* after a plural subject – for English it has the plural number (*lives*), and, for German, the number is singular (*Leben*[2]).

(1) Some people lost their lives, and I'm still alive, so I'm happy. (COCA; 2012)

(2) 13 Personen kamen ums Leben. (DWDS-Kernkorpus1900–1999; 1945)

The present paper takes the first steps towards verifying the supposed differences between English and German by means of a pilot corpus-based study conducted with the use of the Oslo Multilingual Corpus (OMC) – a parallel corpus containing original texts and their translations. Thus, the research question addressed is "Do English and German differ with regard to number preference in objects, adverbials, and modifiers following plural subjects?".

As the point of departure for the present study, let us look at two sentences extracted from the OMC and describing the same event – (3) is an excerpt from an original English text, and (4) its translation into German.

(3) (...) ladies with frozen smiles and swaying crinolines; their wigs were powdered, their cheeks pocked with beauty spots, and there were black bows tied around their necks. (OMC)

(4) (...) Damen mit einem erstarrten Lächeln und schwingenden Krinolinen; ihre Perücken waren gepudert, ihre Wangen mit Schönheitsflecken übersät, und um ihren Hals waren schwarze Schleifen gebunden. (OMC)

As we can see, in both sentences the subject is plural – *ladies* in English, *Damen* in German. In the English version, all the ladies have *frozen smiles* and *swaying crinolines*, *powdered wigs*, and *black bows* tied *around their necks*. So, there is an absolute correspondence between the plural number of the subject and the nouns in the predicate part of the sentence. This is, however, not the case in the German translation. The ladies (plural form *Damen*) in the German version seem all to be flashing one *frozen smile* (*mit einem erstarrten Lächeln*) but wear multiple *crinolines* and *wigs*. Around their singular *neck* (*um den Hals*) plural *black bows*

[1] "Kongruenz im Numerus" (Eng. "Number agreement in German") Duden online. URL: https://www.duden.de/sprachwissen/sprachratgeber/Kongruenz-im-Numerus (accessed on: 01.12.2022).

[2] The singular number is indicated by *s* in *ums Leben*.

(*Schleifen*) are tied. This lack of correspondence is precisely what the current research seeks to explore.

The three nouns looked at in the present study are *head*/*heads*, *life*/*lives*, *voice*/*voices*, and their German equivalents. The investigation is bidirectional – both German translations of English originals and English translations from German are examined. The results show that the preference for the distributive plural is indeed the norm for English as it can be used in almost any context, while in German it is a much more context- and noun-specific feature.

On the methodological side, the study applies a quantitative and qualitative analysis of corpus data. The examples are retrieved from the corpus with the use of the in-build search engine, which is followed by a careful qualitative assessment of results. The items we search for are the three nouns (*head*/*life*/*voice*) in both plural and singular form, preceded by posessive pronouns indicating the plural number of the subject (for English *their*/*our*, for German depending on the gender of the noun, and including definite and indefinite articles) to limit our searches to examples with plural subjects. It is worth noting that due to cross-linguistic differences (e.g., there are four cases in German), the number of corpus searches for German is much greater for each of the nouns than for English[3].

This small-scale corpus-based analysis aims at providing us with an idea of the general tendencies, possible problems and exceptions. These results are further supplemented by results of acceptability ratings which offer more insights into the linguistic intuitions of native speakers of the languages in question. The platform chosen for the acceptability survey is Prolific. It combines good recruitment standards with reasonable costs, clearly informing participants that they are being recruited to participate in a research study (Palan & Schitter, 2018).

References

Aarts, B., Chalker, S. & Weiner, E. (2014): *The Oxford Dictionary of English Grammar* (2nd ed.). Oxford: Oxford University Press.

Palan, S. & Schitter, C. 2018. Prolific.ac - A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 p. 22-27.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985): *A Comprehensive Grammar of the English Language*. Edinburgh: Longman.

The Oslo Multilingual Corpus (1999-2008), the Faculty of Humanities, University of Oslo. The Oslo Multilingual Corpus is a product of the interdisciplinary research project Languages in Contrast (SPRIK)[4], directed by Stig Johansson and Cathrine Fabricius-Hansen, and compiled by the OMC corpus team. http://www.hf.uio.no/ilos/english/services/omc/.

[3] In particular, to illustrate what is meant, the following variants were entered into search engine for the noun *head/heads*: i) for English: *their head*, *our head*, *their heads*, *our heads*; ii) for German: *die Köpfe, der Köpfe, den Köpfen, ihre Köpfe, ihrer Köpfe, ihren Köpfen, unsere Köpfe, unserer Köpfe, unseren Köpfen* (distributive plural); *der Kopf, des Kopfes/Kopfs, dem Kopf, den Kopf, ihr Kopf, ihres Kopfes/Kopfs, ihrem Kopf, ihren Kopf, ein Kopf, eines Kopfes/ Kopfs, einem Kopf, einen Kopf, unser Kopf, unseres Kopfes/Kopfs, unserem Kopf, unseren Kopf.*

[4] Oslo Multilingual Corpus - background and use: http://www.hf.uio.no/ilos/english/services/omc/; accessed on 16.10.2022.

Fake news crossing spaces and language borders: A contrastive analysis of English and Lithuanian

Jurate Ruzaite

This paper aims to examine the language of online fake news in English and Lithuanian using the methods of contrastive corpus linguistics. Contrastive analysis of disinformation features is not as straightforward as it may seem. The present study thus not only reports research results but also addresses some methodological issues encountered in contrastive analysis of fake news.

The existing research on disinformation primarily focuses on comparing fake and factual news. However, systematic cross-linguistic studies are still limited (see Damstra et al. 2021), Humprecht's (2019) comparative analysis of English and German being one of the few cross-linguistic studies available to date. Some studies (Humprecht 2019, Siwakoti et al. 2021) suggest significant cross-country differences in disinformation narratives, but the extent of linguistic differences in fake news remains unclear.

In this study, the texts used for the empirical analysis were identified as fake news through fact-checking, and some of the fake facts had already been debunked by mainstream media. The data includes 40 Lithuanian disinformation texts about COVID-19 (18,111 tokens) obtained from the propaganda portal *minfo.lt*. When compiling the Lithuanian corpus, it appeared that 32 texts (80%) were based on English-language sources. To examine how the English texts were exploited for Lithuanian disinformation, the original English texts were collected for comparative analysis (32 texts; 52,264 tokens).

The analysis addresses the following research questions: (1) What are the linguistic properties of Lithuanian and English fake news?; (2) What is the lexical diversity (type-token ratio, TTR) of the texts?; (3) How are emphatics and tentative language used in Lithuanian and English data? These RQs stem from earlier research indicating that low lexical diversity, emphatics, and tentative language are strongly associated with disinformation (e.g. Ribeiro Bezerra 2021). Prior research suggests that fake news is marked by simplicity, and thus TTR in disinformation texts is lower than in factual texts (Kumar & Vardhan 2021). Emphatics and tentative language are focused on as they serve as useful indicators of bias: tentative language (e.g. *just, only*) is a category of epistemological bias, and emphatics (e.g. *very, absolutely*) are a category of framing bias (Recasens et al. 2013). In addition, emphatics are associated with sensationalism and are used as part of propaganda techniques aiming to persuade users at an emotional rather than cognitive level (Damstra et al. 2021, Staender et al. 2021).

The results have important implications regarding several major issues. First, the very size of disinformation corpora is of a major concern. It is relatively uncomplicated to compile a corpus of disinformation texts in English; however, when lesser used languages like Lithuanian are examined, the number of texts available is highly restricted.

A second important issue is the limited amount of original content in Lithuanian disinformation. Since most of the Lithuanian content comes from English sources, it is questionable whether it is likely that some distinct language-specific features can emerge in Lithuanian fake news. When analysing such fluid phenomena as disinformation, the role of language arguably becomes just instrumental: English texts are generated fast and profusely, can be easily translated by fake news producers into different languages and disseminated in fluid media spaces, often making it impossible for the researcher to know the original source, the translation tool, or the author's personal input into the output text. When large amounts of data are collected for corpus analysis, this meta-information can be difficult to control, but in contrastive linguistics these are central variables.

In terms of the content modifications in the Lithuanian texts, they range from very close translations to highly abridged summaries of the source texts. A general trend is that disinformation texts are shorter in Lithuanian. Typically, the parts of the source text that are omitted are the list of references, section titles, and some passages containing supposedly unnecessary information.

Regarding the linguistic properties, the type-token ratio (TTR) is very low in English data (0.14) but considerably higher in Lithuanian (0.39), which could be a result of some typological differences between the two languages. Intensifiers are almost equally distributed in both datasets, but tentative language is of a considerably higher incidence in English. Such trends suggest that the language of 'fake news' tends to be simple; however, Lithuanian fake news aim at sensationalism by retaining the same frequency of emphatic wording but reducing the tentative tone.

References

Damstra, Alyt, Hajo G. Boomgaarden, Elena Broda, Elina Lindgren, Jesper Strömbäck, Yariv Tsfati & Rens Vliegenthart (2021) What Does Fake Look Like? *A Review of the Literature on Intentional Deception in the News and on Social Media, Journalism Studies*, 22:14, 1947-1963, DOI: 10.1080/1461670X.2021.1979423

Humprecht, E. (2019). Where 'Fake News' Flourishes: A Comparison Across Four Western Democracies. *Information, Communication & Society* 22:13, 1973–1988.

Kumar, B. Tirupathi & Vardhan, B. Vishnu. (2021). A Stylistic Feature Based Approach for Fake News Spreaders Detection. *Journal of Tianjin University Science and Technology*, 54:9, 190–209. DOI 10.17605/OSF.IO/Z9DW5

Recasens, Marta, Danescu-Niculescu-Mizil, Cristian & Jurafsky, Dan. (2013). Linguistic Models for Analyzing and Detecting Biased Language. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1650–1659, Sofia, Bulgaria, August 4-9, 2013.

Ribeiro Bezerra, Jose Fabio (2021). Content-based fake news classification through modified voting ensemble. *Journal of Information and Telecommunication*, 5:4, 499-513, DOI: 10.1080/24751839.2021.1963912

Siwakoti, Samikshya, Kamya Yadav, Isra Thange, Nicola Bariletto, Luca Zanotti, Alaa Ghoneim, and Jacob N. Shapiro. (2021). Localized Misinformation in a Global Pandemic: Report on COVID-19 Narratives around the World. *Empirical Study of Conflict, Princeton University*, pp. 1-68, March 2021. https://esoc.princeton.edu/publications/localized-misinformation-global-pandemic-report-covid-19-narratives-around-world.

Staender, Anna, Edda Humprecht, Frank Esser, Sophie Morosoli & Peter Van Aelst. (2021). Is Sensationalist Disinformation More Effective? Three Facilitating Factors at the National, Individual, and Situational Level. *Digital Journalism*, 10:6, 976–996, DOI:10.1080/21670811.2021.196631

Titles of Research Papers Revisited: a Cross-linguistic and Cross-disciplinary Analysis

Jolanta Šinkūnienė

One of the essential aspects of research articles are titles, which, despite their inherent brevity, are considered to be "serious stuff" (Swales 1990: 224). This is not surprising, as titles have major significance in all stages of an article's 'life', i.e. pre-submission, post-submission and post-publication, because they determine to a large extent the focus of the author and the interest of editors, reviewers and readers (Eva 2013: 33). There is quite an extensive body of research into titles of research papers which has delved into the length and punctuation of titles (Lewison & Hartley 2005; Hartley 2007), their syntactic and structural features (Soler 2007; Wang & Bai 2007), pragmatic intentions of the authors (Haggan 2004) or a combination of all these features across a range of science fields and from a diachronic perspective (Hyland & Zhou 2022; Jiang & Hyland 2022). The results of these studies reveal clear-cut disciplinary differences largely determined by epistemologies of different science fields, as well as observable changes over time, one of which is a growing desire of the authors for a greater engagement with readers (Jiang & Hyland 2022). This is again hardly surprising since with an exponential growth of research production, scientific authors have to be competitive and promotional, in a sense "marketing their own research" (Breivega et al. 2002: 220), which apparently can be achieved with the help of a successful title.

There is a growing body of research into titles in English, however, research paper titles in languages other than English are very scarcely researched. Yet they constitute an interesting field since it is important to compare academic writing patterns and traditions not only across disciplines, but also across languages, especially 'small' languages, which do not have such an extensive research production as English. The aim of this paper is to compare the features of research article titles in a number of disciplines of social sciences and humanities in English and Lithuanian. The total number of research article titles under analysis is nearly 1,000. Following methodology devised by Hyland and Zhou (2022) and Jiang and Hyland (2022), the analysis looks into the length of titles, their syntactic structure, and their focus. The paper also employs an ethnographic approach as it provides insights from scholars in the analysed disciplines regarding their own practice of title construction and their own reaction to the titles of research articles published by their peers. The results show that scholars writing in English tend to employ compound titles to a larger extent in nearly all analysed disciplines in comparison to Lithuanian scholars. This strategy allows to highlight interesting angles of the study, play with words or make intertextual references. The analysed articles in English probably rely on these strategies more since they are aimed at a much larger audience and it is important to make them stand out in order to attract the reader. This is especially evident in the English article titles in the disciplines of law and sociology. Interrogative titles were not frequent in both languages with the exception of law article titles in English and sociology article titles in Lithuanian.

References

Breivega, K.R., Dahl, T. & Fløttum, K. 2002. Traces of self and others in research articles. A comparative pilot study of English, French and Norwegian research articles in medicine, economics and linguistics. *International journal of applied linguistics* 12(2): 218–239.

Eva, KW. 2013. Titles, abstracts and authors. In *How to Write a Paper*, ed. by Hall G.M. 5th edn. Oxford: John Wiley & Sons, 33–41.

Haggan, M. 2004. Research paper titles in literature, linguistics and science: dimensions of attraction. *Journal of Pragmatics* 36: 293–317.

Hartley, J. 2007. Planning that title: Practices and preferences for titles with colons in academic articles. *Library & Information Science Research* 29: 553–568.

Hyland, K. & Zou, H. 2022. Titles in research articles. *Journal of English for Academic Purposes* 56.

Jiang, F. & Hyland, K. 2022. Titles in research articles: changes across time and discipline. *Learned Publishing*, 1–10.

Lewison, G. & Hartley, J. 2005. Numbers of words and the presence of colons. *Scientometrics*, 63, 341–356.

Soler, V. 2007. Writing titles in science: An exploratory study. *English for Specific Purposes* 26: 90–102.

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

Wang, Y. & Bai, Y. 2007. A corpus-based syntactic study of medical research article titles. *System* 35: 388–399.

Parallel sessions

Apparently – a corpus analysis of a recent change in spoken English

Karin Aijmer

The present study was inspired by the observation that the adverb apparently has increased its frequency over a relatively short time in present-day English. In a sample of five million words from the Spoken BNC2014 (Love et al. 2017) apparently had a frequency of 181.08 occurrences pmw to be compared with 83.55 tokens pmw in (the demographic component of the) BNC1994. My aim is to investigate the changes reflected in the rising number of tokens of apparently by comparing the frequency and uses of apparently in the two corpora. The rapid surge of frequency also leads to questions about the age and gender of the innovators and about the factors motivating the changes. According to Chafe, apparently is an evidential inferential adverb signalling a conclusion based on sensory evidence or hearsay (Chafe 1986). A characteristic feature of the evidential apparently is that it can also have an epistemic component conveying the speaker's weak commitment ('it is probable that'). Moreover, apparently has acquired an indirect-reportative evidential meaning involving access to generalized second-hand information which is potentially unavailable to the speaker or hearer (Marín Arrese 2017). The preliminary quantitative results (based on a pilot study) indicate that in both BNC 1994 and the Spoken BNC2014 apparently had indirect-reportative meaning except in a few occurrences where the evidence was explicitly referred to. The corpus findings also show that apparently in the Spoken BNC2014 is often used in interactional contexts where its evidential meaning to refer to the source of information has been weakened. It is argued that in these contexts the use of apparently can convey a mirative overtone that the information is unexpected in relation to the speaker and the hearer (De Lancey 1997). The mirative interpretation is particularly clear when the subject of the verb in the clause introduced by apparently is the first person:

(1) and apparently I told her I was like yeah there 's someone standing next to me

The changes undergone by apparently leads to question about the age and gender of the innovators and their social goals. The sociolinguistic analysis shows that most of the speakers using apparently are young and that there is a rise in the proportion of female speakers using apparently in the Spoken BNC2014. The close association between the frequency of apparently and young female speakers further suggests that these speakers use apparently for identifying themselves as members of a social group.

Understanding corpus text prototypicality: A multifaceted problem

Laurence Anthony, Nicholas Smith, Sebastian Hoffmann and Paul Rayson

Prototypicality is a complex, multifaceted concept relating to the centrality and typicality of examples in a category. While prominent in cognitive psychology and linguistics, it is often overlooked in corpus studies. Corpora are ideally built to be representative of a target domain or language variety. To achieve this goal, corpus builders need to identify an accurate sampling frame and collect relevant texts that capture the diversity of language in and across the sampling categories. In practice, however, corpora are built within the limitations of text availability, time, and human resources leading to questions about the *suitability / prototypicality* of individual texts in a corpus and their effect on the representativeness of the corpus as whole. Prototypicality also comes into play at the analysis stage. Most corpus analysis approaches use the corpus as a whole as the unit of analysis, including concordance and keyword analysis. To validate findings, a necessary but often omitted step is the close reading of individual texts. Here, a significant challenge is identifying which texts to read. A researcher may decide to randomly choose texts, but it is an open question if such texts are *representative/prototypical* of the corpus. Prototypicality also comes into play when corpora are used for pedagogic purposes, such as Data-Driven Learning (DDL). In these situations, there is often an implicit conflation of two facets of prototypicality, namely frequency of use and closeness to an ideal, particularly in the case of expert writing.

In this paper, we first outline the multifaceted character of corpus text prototypicality. Next, we describe experiments that attempt to rank the prototypicality of individual corpus texts at different linguistic levels as a guide to choosing texts for close reading or excluding texts from a corpus at the data collection stage. Results using a modified version of the *ProtAnt* tool (Anthony and Baker, 2015) show prototypicality rankings can be dramatically affected by the linguistic level of analysis applied. Standard keywords effectively rank the prototypicality of texts in terms of topic, but the results can be enhanced using key semantic tags. On the other hand, key part-of-speech (POS) tags allow for a more nuanced view of text prototypicality centered on stylistics. The results also reveal the limitations of current corpus software tools and offer suggestions for how new tools might be developed to increase our understanding of prototypicality at the textual level.

Corpus-based investigation of ambiguity resolution in scientific and academic writing

John Blake

This paper presents the results of a corpus investigation of ambiguity in scientific and academic writing. In natural language processing, ambiguity resolution is an unsolved problem that is central to increasing parsing accuracy. However, from the standpoint of increasing clarity in scientific and academic writing, ambiguity is also a central concern. Readers of research articles and academic essays frequently notice expressions that are vague or ambiguous. Vague expressions may be avoided by selecting more specific terminology or through elaboration. Ambiguity in writing may arise unintentionally causing confusion in the reader. Causes of ambiguity include when there is more than one contextually possible meaning of a word, when it is unclear which expressions are being modified, and when there are multiple candidates for the anaphoric referent of pronouns. Scientific and academic texts often display high degrees of nominalization. This conformation to the generic integrity of formal writing enables concepts to be abstracted and discussed. The transformation of sentences to clauses, clauses to phrases, and phrases to words increases concision, but potentially at the expense of clarity. However, such transformations may result in ambiguity, particularly when the grammatical relation in compressed noun phrases needs to be deduced.

A specialist corpus of 200 draft research articles from the fields of information, materials and knowledge science was compiled from online submissions to a university writing center. Most papers were jointly authored, with the main author being a postgraduate student in either a master's or doctoral program. Instances of ambiguity were manually identified, classified and annotated as referential, syntactic or lexical using the UAM Corpus Tool. For each identified instance, ways to resolve the ambiguity were considered by the annotator.

Manual analysis revealed numerous instances of referential, syntactic and lexical ambiguity. Referential ambiguity frequently resulted when pronouns, especially it and this, were used anaphorically. Syntactic ambiguity often occurred when relative clauses and prepositional phrases were used following coordinating conjunctions (e.g. and). Lexical ambiguity was attributed to words with multiple meanings (e.g. get), confusing conjunctions (e.g. inclusive or exclusive or) and words with preferred meanings (e.g. significant) or frames (e.g. ditransitive or monotransitive).

The paper concludes with practical ways to resolve the three types of ambiguity through grammatical and lexical choice.

Busy to develop: Another English progressive construction

Adri Breed and Frank Brisard

Breed (2022) critically evaluates several claims made in the linguistic literature about the use of "busy" in South-African English (SAfrE) as discussed in various sources (see a.o. Branford 1980, Bekker 2012, Bowerman 2004, Lass 1995, Siemund 2019, Lass and Wright 1986, and DOSAE 1996). One of the claims she evaluates is that instances of the construction [BUSYPROG XCOMP VINF], such as "I'm busy to develop a course," which are sometimes present in SAfrE corpus data, should be considered nonstandard and attributed to ESL speakers with Afrikaans as L1 translating the Afrikaans *busy*-progressive form, such as "Ek is besig om 'n kursus te ontwikkel" (see Mesthrie 1999). Breed (2022) argues that this construction is also found in corpora of other varieties of English, such as Tanzanian English and Bangladeshian English, and should therefore be considered an (if marginally) available construction in English. Evidence for this argument is provided through providing examples from historical corpus data, which shows that this construction was in use as far back as the 1500s.

This presentation extends Breed's (2022) study by discussing the results of a corpus analysis of the [BUSYPROG XCOMP VINF] construction in English from both a historical and a contemporary perspective. Historical corpora, including EEBO (1470s-1690s), EMMA (1623–1757), and COHA (1820-2019), are analysed to demonstrate the development of the construction as a progressive form and its evolution over time. This evolution is set against the emergence of the English [V-ingPROG] construction between the 17th and 18th century (see Smith 2007 and Kranich 2010). Recent corpora of various varieties of English, such as the GloWbE, NOW, and iWeb corpora (Davies 2023), are used to compare the [BUSYPROG XCOMP VINF] construction with its competing form, the more dominant English [V-ingPROG] construction. The analysis primarily focuses on frequency differences, subject and predicate collocation differences, and the language use contexts (such as register and genre).

References

Bekker, Ian. 2012. The story of South African English: A brief linguistic overview. *International Journal of Language, Translation and Intercultural Communication* 1. https://doi.org/10.12681/ijltic.16 (accessed on 4 January 2022). 139–150.

Bowerman, Sean. 2004. White South African English: Morphology and syntax. In Rajend Mesthrie (ed.), *Varieties of English: Africa, South and Southeast Asia*, 472-484). Berlin: Walter de Gruyter.

Branford, Jean. 1980. *A dictionary of South African English.* https://archive.org/details/dictionaryofsout00bran (accessed on 3 December 2021).

Breed, Adri. 2022. The Dutch and Afrikaans "BUSY progressive" is busy to exist in English. Paper presented at the 'A Germanic Sandwich 2022' conference at The Institute of Dutch Language and Literature, University of Cologne, Germany. [Online] 17-18 March 2022.

Davies, Mark. 2023. English-Corpora.org. https://www.english-corpora.org. (Date of access: 7 February 2023).

Lass, Roger & Susan Wright. 1986. Endogeny vs. Contact: 'Afrikaans influence' on South African English. *English World-Wide* 7(2), 201-223.

Lass, Roger. 1995. South African English. In Rajend Mesthrie (ed.), *Language and Social History: Studies in South African Sociolinguistics*, 89-106. Cape Town: David Philip Publishers.

Siemund, Peter. 2019. Regional varieties of English: non-standard grammatical features. BasAarts, Jillian Bowie and Gergana Popova (eds.), *Oxford Handbook of English Grammar,* 604-629. Oxford: Oxford University Press.

Mesthrie, Rajend. 1999. Syntactic change in progress: Semi-auxiliary busy in South African English. *University of Pennsylvania Working Papers in Linguistics* 6(2), Article 6. https://repository.upenn.edu/pwpl/vol6/iss2/6 (accessed on 30 November 2021).

Smith, K. Aaron. 2007. The Development of the English Progressive. Journal of Germanic Linguistics, 19(3)

Kranich, Svenja. 2010. The Progressive in Modern English: A Corpus-Based Study of Grammaticalization and Related Changes. Leiden: Brill.

Coming to grips with *rather* elusive adverbs: On EN *rather*, DU *eerder* and FR *plutôt*

Lieselotte Brems and *Lobke Ghesquiere*

As argued in Brems & Ghesquière (2017) and Brems et al. (2020), the English adverbial markers *rather (than)* and its Dutch equivalent *eerder (dan)* can both express preference (1) or contrast (2), and intensity (3). In addition, Dutch *eerder (dan)* has temporal uses as in (3), which are obsolete for English *rather (than).*

(1) EN *Although I wholeheartedly support this aim, I would much rather see such aid eliminated altogether.*[5]

DU *Hoewel ik deze doelstelling met hart en ziel ondersteun, zou ik nog veel liever zien dat er helemaal een einde aan kwam.*

(2) EN *Competition between the regions will certainly strengthen rather than weaken the European Union.*

DU *Concurrentie tussen de regio's zal de Europese Unie versterken, niet verzwakken.*

(3) DU *Hoe eerder wij deze afronden des te beter*

EN *The sooner we achieve this, the better it will be for all of us.*

Although past monolingual and contrastive study of these adverbs has already gone some way in disentangling the different language-internal uses of these adverbs as well as map out crosslinguistic morphosyntactic and pragmatic-semantic differences and similarities, the translation study aims to further contribute to our understanding of these adverbs. It specifically aims to account for contextual clues triggering a certain reading, and hence translation, of the adverbs concerned. As such, this paper aims to fine-tune and complement earlier work and find further proof for the relevance of using parallel corpora and translation data for the study of semantic shifts, as argued for in Beeching (2013) f.i.

The sentence-aligned Dutch and English subcorpora of the Europarl parallel corpus, accessed throughout the SketchEngine, are used to compare original Dutch and English data on *eerder* and *rather* with English and Dutch translations respectively. We will draw up typologies of the different uses of *rather (than)* and *eerder (dan)*, comparing them and assessing their degree of intertranslatability by characterizing them semantically and structurally, and then quantify the results.

.

References

Beeching, K. (2013) A parallel corpus approach to investigating semantic change. In *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson*, Aijmer, K. and B.Altenberg (eds.) Amsterdam: John Benjamins. 103-125.

Brems, L. & L. Ghesquière. 2017. Time, preference and intensity: A contrastive study of rather (than) and eerder (dan). Paper presented at ICAME

Brems, L. L. Ghesquière & G. Vanderbauwhede. 2020. A rather interesting topic: A contrastive study of English rather, Dutch eerder and French plutôt. In Translating and Comparing

[5] All examples taken from Europarl.

Languages: Corpus-based Insights Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference, S. Granger & M-A Lefer (eds.). UPL: 215-236.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit* 5: 79-86

Traugott, E.C. & E. König. 1991. The semantics-pragmatics of grammaticalization revisited. In E.C. Traugott & B. Heine (eds.), *Approaches to Grammaticalization*, Amsterdam: Benjamins. 189–218.

Complex premodifiers in World Englishes: Comparing evolutionary and contact-induced explanations

Marcus Callies and Turo Vartiainen

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). In American English, this trend has also been observed for complex premodifiers, such as the A-to-V construction with tough-predicates (e.g. a difficult-to-answer question) and the comparative than-construction (e.g. a larger-than-life character; Günther 2018). Considering the global influence of American English (Mair 2013; Gonçalves et al. 2018; Schneider 2020), complex prenominal modifiers most likely have spread to World Englishes, but their diffusion patterns have received little attention in previous research (however, see Mazaud 2004). In this paper, we examine three complex premodifier constructions in 20 varieties of English by using the corpus of Global Web-based English (GloWbE; Davies 2013). These constructions include the complex prepositional construction (1), the comparative than-construction (2), and the prenominal tough-construction (3).

(1) His strength was the combination of off-the-charts vision and exacting precision. (GloWbE, CA)

(2) What's this sleek, sexy Tokyo surprise doing in the less-than-trendy area of Sa Ying Pun? (GloWbE, HK)

(3) Below is my very simplified and easy-to-understand guide […]. (GloWbE, KE)

We examine the frequencies of these constructions in light of two explanatory factors intended to account for differences in NP-complexity in African and Asian varieties of English: i) SLA-induced structural simplification and ii) language contact. As for i), ESL varieties of English are assumed to exhibit simpler constructions than their input varieties due to general cognitive processes of SLA with the extent of simplification depending on the variety's evolutionary progress in Schneider's (2007) Dynamic Model. As for ii), the typological features of the major substrate languages are expected to affect the overall incidence of head-final/head-initial NPs and the degree of complexity of pre- and postmodifying structures in the respective variety (e.g. Schilk & Schaub 2016; Brunner 2017; Akinlotan 2018; Brato 2020).

Our findings suggest that both factors are relevant and partially work together: in our data, complex premodifiers are most frequently attested in several Southeast Asian Englishes that are fairly advanced in Schneider's Dynamic Model and whose main substrate languages have head-final syntax. By contrast, less advanced varieties with head-final substrate languages (e.g. Sri Lanka) and advanced ones with head-initial substrate languages (e.g. West African Englishes) exhibit the lowest frequencies of complex premodifiers in the data.

Building consensus on climate change: the language of diplomacy in media interviews

Silvia Cavalieri, Sara Corrizzato, Valeria Franceschi

A great deal of research has been devoted to the language used in specific sectors, such as politics, science, business or medicine and the various ways through which interlocutors influence communication worldwide; yet the linguistic strategies used by diplomats to convey or negotiate meaning and values are not fully conceptualized and discussed in linguistics. With careless language, experts in diplomacy are unable to achieve their goals. Their feelings and opinions are determined and demonstrated not only by what they say and how they say it, but also by what remains unsaid and left to the interpretation of their message recipients. Considered "as a special way of expressing the subtle needs of the diplomatic profession" (Nick 2001), the language employed within this field is of central relevance to understand how institutional representatives (politicians, diplomats, field experts, leading members of organizations) address very sensitive topics, such as climate change and its consequences, within an international panorama. This study presents a mixed-methods approach which includes both quantitative and qualitative analysis of a corpus of media interviews to leading figures discussing the social, economic and environmental consequences of climate change. Gathered between 2021 and 2022 from major English-language international broadcasting companies, the corpus (i.e. the Interdiplo corpus, an ongoing project at the Dept. of Foreign Languages and Literatures – University of Verona, Italy) includes both native and non-native speakers that use English as a vehicle for communication. In this study the linguistic and pragmatic strategies of subjectivation and depersonalization (Fraser 2009; Martin-Martin 2008; Hyland 1998) will be investigated to understand how and to what extent speakers choose (in)directness to avoid any conflict-related situation. Preliminary findings demonstrate that, on the one hand, journalists' discursive strategies evolve from neutral to more confrontational formulations to challenge interviewees when dealing with hot issues such as climate change; on the other hand, data show that diplomats' responses tend to mitigate conflict by showing empathy and promoting a win-win approach to maximize results as well as to seek consensus and facilitate dialogue on climate change.

References

Fraser, B. (2009). "An Account of Discourse Markers." *International review of Pragmatics*, 1(2): 293-320.

Hyland, K. (1998). *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins Publishing.

Martin-Martin, P. (2008). "The Mitigation of Scientific Claims in Research Paper: A Comparative Study." *IJES*, 8(2): 133-152.

Nick, S. (2001). "Use of Language in Diplomacy." In J. Kurbalija and H. Slavik (eds.) *Language and Diplomacy*. Malta: DiploProjects. 17-21.

That/zero Complementisers in some Varieties of English

Miriam Criado-Peña

The present paper analyses the distribution of *that*/zero as object clause connectives in some varieties of English. An object clause is defined as “a type of dependent clause used to complete the meaning relationship of an associated verb or adjective in a higher clause” (Biber et al. 1999: 658). In English, the most common type is introduced by the complementiser *that* (e.g., *I know that she is coming*), which is sometimes omitted leaving an asyndetic zero-clause (e.g., *I know she is coming*) (Conde-Silvestre and Calle-Martín 2015: 58). The competition of these two complementisers has received some scholarly attention in British and American English from both a synchronic (Elsness 1984; Biber et al. 1999) and a diachronic perspective (Fanego 1990a-b; López-□ouso 1996; Suárez-Gómez 2000; □alle-Martín and Romero-Barranco 2014; □onde-Silvestre and Calle-Martín 2015), while their distribution in the so-called New Englishes has been generally left unexplored, with the exception of Kruger and Van Rooy (2019) and Van Rooy (2021), who investigated the phenomenon in South African Englishes. In light of this gap in the literature, this work explores the variation in the choice of complementiser as regards *that*- and zero-clauses in three Asian varieties of English (i.e., Indian English, Hong Kong English and Philippine English), taking their superstrate varieties (i.e., British English and American English) as points of departure, with the following objectives: a) to analyse the distribution of *that*/zero across varieties, and across speech and writing; b) to ascertain the participation of a series of linguistic and extra-linguistic conditioning factors; and c) to determine whether the preference for *that*/zero is more related to the influence of their superstrates or the transference of substrate features. For the purpose, the Indian, Hong Kong and Philippines components of the *International Corpus of English* (ICE) have been used as source of evidence along with the *British National Corpus* (BNC) and the *Corpus of Contemporary American English* (COCA). Preliminary results indicate that the choice of complementiser in Asian Englishes is primarily derived from the transference of substrate features rather than from the introduction of superstrate ones.

References

Biber, D., et al. (1999). *Longman grammar of spoken and written English*. London: Longman.

Calle-Martín, J. & Romero-Barranco, J. (2014). On the use of that/zero as object clause links in early English medical writing. *Studia Neophilologica, 86(1), 1–16.*

Conde-Silvestre, J. C. & Calle-Martín, J. (2015). Zero *that*-clauses in the history of English: A historical sociolinguistic approach (1424–1681). *Journal of Historical Sociolinguistics*, *1*(1), 57–86.

Elsness, J. (1984). That or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies*, *65*(6), 519–533.

Fanego, T. (1990). Finite complement clauses in Shakespeare’s English. *Studia Neophilologica*, *62*, 3–21 & 129–149.

Kruger, H. & Van Rooy, B. (2019). A multifactorial analysis of contact-induced change in speech reporting in written White South African English (WSAfE). *English Language and Linguistics*, *24*(1), 179–209.

López-Couso, M. J. (1996). A look at *that*/zero variation in Restoration English. In D. Britton (Ed.), *English historical linguistics 1994: Papers from the 8th international conference on historical linguistics* (pp. 271–286). Amsterdam & Philadelphia: John Benjamins.

Suárez-Gómez, C. (2000). That/zero variation in private letters and drama (1420–1710): A corpus-based approach. *Miscelánea. A Journal of English and American Studies*, *21*, 179–204.

Van Rooy, B. (2021). Grammatical change in South African Englishes. *World Englishes*, *40*(1), 24–37.

Comparing English to languages other than English: Addressing challenges in comparability to investigate global discourses

Niall Curry

Research in contrastive linguistics has long critiqued the comparability of the English language with languages other than English, labelling the former as incommensurable, owing to its roles as an international lingua franca and the language of modern science. Nevertheless, many studies draw on comparisons of English with other languages to unpack and understand the shared and disparate practices embodied in comparable texts across languages. This is typically done with a view to understanding what these texts tell us about the communities that produce and engage with said texts. To conduct such a study effectively, issues of comparability must be investigated in order to ensure that like with like is being compared when comparing across languages.

Recognising this view, this paper contributes to the growing body of research on the discourses of climate change, which occupies an important niche in applied linguistics literature. Work in this area has revealed that cultural variation plays a critical role in determining how issues such as deforestation and climate scepticism, for example, appear to be of varying importance across cultures and languages. Seeing the English language as the global language of science, this paper presents a corpus-based contrastive analysis of English, French, and Spanish that interrogates the inherent comparability of climate crisis discourses and compares and contrasts academic and parascientific texts on the climate crisis. This is done with a view to understanding how variation in climate change discourse is socially constructed across disciplines, genres, and languages. Drawing on a range of corpus techniques, including key word analysis and function-to-form/form-to-function analysis, the aim of this paper is to show correspondences and differences in climate change discourses across the texts studied with a view to identifying the varying ways in which climate change knowledge is constructed. In doing so, the study identifies potential points of ambiguity and confusion in the linguistic construction of climate change when contrasted on the global and multidisciplinary stage.

The results of this study offer a number of key contributions to corpus linguistic research on climate change. First, they add a much needed multilingual perspective on a global issue that has largely been studied in the English language. Second, the results illustrate key differences that occur in the social construction of climate change knowledge across language, discipline, and genre. The identified differences point to ways in which the linguistic construction of climate change knowledge can contribute to misconceptions, misrepresentations, and confusions surrounding this issue, as information and knowledge developed in the English language move across cultural and linguistics spaces. This movement, in turn, can hamper the development of a coherent, collective understanding of climate change as a global, multicultural, and transdisciplinary issue.

Advice and uptake in the Clinton Email Corpus

Rachele De Felice

In a recent publication, Põldvere et al. (2022) set out a framework for identifying and describing advice-giving constructions in English and the types of uptake and responses they elicit. Although this framework is detailed and comprehensive, it is based only on data from corpora of spoken British English (the London-Lund Corpora, LLC). It is therefore useful to assess its wider applicability by testing whether its findings can be generalised to different linguistic spaces. This paper does so by assessing the framework against the Clinton Email Corpus (CEC), which differs from LLC data in two dimensions or linguistic spaces: geographical variety (American vs. British) and medium of communication (written email vs. spoken language).

More specifically, the following research questions are addressed. 1) Do the interlocutors of the CEC use the same forms of advice-giving as those identified in the LLC? 2) Where uptake (in the form of email replies) is available, are the same patterns observed? 3) Are there any differences in advice-giving between interlocutors of different seniority and/or different levels of familiarity?

We use a subset of the CEC consisting of 500 emails where the relationships between interlocutors are known (De Felice and Garretson 2018). This preserves information about hierarchical disparities between interlocutors, mirroring a key aspect of the Põldvere et al. 2022 study, which also investigates the role of relationships in the advising interaction.

We search this dataset for all the advice constructions listed in Põldvere et al. (available from https://www.cambridge.org/files/8516/6444/7762/Annotation_Manual.pdf), resulting in around 1500 instances of advice-giving. One of the most striking differences observed so far is the much higher frequency of strong directives and unambiguous forms of advice in the CEC compared to the LLC (e.g. *we must/have to, I recommend, my advice is…*). Further qualitative analysis will determine whether this is influenced by the different linguistic variety or the communicative setting, and how uptake varies depending on the advice construction.

The small size of the corpus also allows for further manual inspection to identify any advice-giving exchanges not captured by the initial list of search terms, ensuring further validation – or expansion – of the original study's coverage.

References

Rachele De Felice and Gregory Garretson (2018). Politeness at work in the Clinton Email Corpus: a first look at the effects of status and gender. *Corpus Pragmatics* 2 (3), 221-242.

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

Fine-tuning the general linguistic range descriptor at CEFR B1 and B2 levels: Noun Phrase Complexity, Text Type and Variety

María Belén Díez-Bedmar and Mark Brenchley

A common criticism of the CEFR and its Companion Volume (Council of Europe, 2001, 2020) centres on the generic nature of the descriptors, which it is argued can act as a barrier to classroom use (Figueras, 2007; Hulstijn, 2007; Little, 2007; North, 2007; Díez-Bedmar, 2018; Díez-Bedmar and Byram, 2018). This paper provides corpus-based evidence for this critique, arguing the value of supplementing the CEFR with finer-tuned and, crucially, genre-based descriptors.

We present a genre-based analysis of noun phrase (NP) complexity (Biber et al.,

2021; Díez-Bedmar & Pérez-Paredes, 2020; Durrant, Brenchley, & McCallum, 2021; Durrant & Brenchley, 2022) as exhibited in two subcorpora in the FineDesc corpus, a novel collection of B1/B2/C1 level CertAcles exam writing by Spanish learners of English. Learner writing is contrasted with a contemporary L1 reference corpus, the BNC2014 Baby+ (Brezina, Hawtin & McEnery, 2021), to yield a 6-way set of text types grounded in two genres (emails, online prose) produced both by learners at two CEFR levels (L1 Spanish EFL writing at B1 - 84 texts, 13154 words - and B2 - 75 texts, 15381 words) and L1 English writers.

From these texts, we extracted the 10 most frequent NP heads, double-coding them according to the type and range of complexity exhibited, using a bespoke taxonomy originally devised for secondary school-level writing (Díez-Bedmar & Pérez-Paredes, 2020) but revised to meet the demands of NP complexity in B1 and B2 writing. The annotated NPs were then subject to non-parametric analyses to identify statistically significant differences across genres and levels.

The result was a nuanced set of patterns, interacting along three dimensions: NP complexity-type, proficiency profile and, crucially, genre. Thus, for example, whilst the online prose exhibited group-level variation regarding the extent of both premodification and postmodification, with 12 significant differences evident, the emails evidenced variation only in postmodification and with only 3 significant differences evidence. Of equal interest was the finding that group-level variation was not consistent with a requirement for B1 writers to develop a capacity to produce greater levels of NP complexity to mirror L1 norms. Whilst this was characteristic of online prose, the blogs pointed to B1 writers also needing to reduce the level of complexity displayed for certain postmodification types.

Our talk will summarise and exemplify these patterns in detail, showing how they provide support for complementing the CEFR with additional sets of more granular, genre-based descriptors informed by L1 and L2 performance.

References

Biber, D., Johansson, S., Leech, G. N., Conrad, S., Finegan, E. (2021). *Grammar of spoken and written English.* Amsterdam: John Benjamins.

Brezina, V., Hawtin, A. & McEnery, T. (2021). The Written British National Corpus 2014 – design and comparability. *Text & Talk - An Interdisciplinary Journal of Language Discourse Communication Studies*, 41, 5-6.

Council of Europe (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2020). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume*. Strasbourg: Council of Europe Publishing.

Durrant, P. & Brenchley, M. (2022) Development of Noun Phrase Complexity Across Genres in Children's Writing, *Applied Linguistics*, amac032. Available online at: https://doi.org/10.1093/applin/amac032

Durrant, P., Brenchley, M., & McCallum, L. (2021). *Understanding development and proficiency in writing: Quantitative corpus linguistic approaches*. Cambridge: Cambridge University Press.

Díez-Bedmar, M. B. (2018). Fine-tuning descriptors for CEFR B1 level: insights from learner corpora. *ELT Journal*, 72(2), 199-209. https://doi.org/10.1093/elt/ccx052

Díez-Bedmar, M. B., & Byram, M. (2019). The Current Influence of the CEFR in Secondary Education: Teachers' Perceptions. *Language, Culture and Curriculum*, 32, 1-15.

https://doi.org/10.1080/07908318.2018.1493492

Díez-Bedmar, M. B. & Pérez-Paredes, P. (2020). Noun phrase complexity in young Spanish EFL learners' writing. Complementing syntactic complexity indices with corpus-driven analyses. *International Journal of Corpus Linguistics*, 25(1), 4-35. https://doi.org/10.1075/ijcl.17058.die

Figueras, N. (2007). The impact of the CEFR. *ELT Journal*, 66(4), 477-485. https://doi.org/10.1093/elt/ccs037

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, 91(4): 663–7. https://doi.org/10.1111/j.1540-4781.2007.00627_5.x

Little, D. (2007). The Common European Framework of Reference for Languages: perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91(4), 645–55. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x

North, B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal*, 91(4), 656–9. https://doi.org/10.1111/j.1540-4781.2007.00627_3.x

Does task type impact EFL accuracy profiles? Insights from the FineDesc learner corpus

María Belén Díez-Bedmar and Jennifer Thewissen

This study is carried out within the framework of the Spanish FineDesc project, funded by the Spanish Ministry of Science and Innovation (PID2020-117041GA-I00 funded by MCIN/AEI/10.13039/501100011033). Taking as a point of departure the Common European Framework descriptors (Council of Europe, 2001, 2020), this project aims to empirically describe the performance of Spanish-speaking learners of English in multiple areas of L2 proficiency (e.g. pragmatics, accuracy, complexity, etc.). To this end, a new learner corpus, the FineDesc learner corpus (Díez-Bedmar, 2022a), is being compiled. It comprises L1 Spanish EFL texts which have been rated by two independent experts in the high-stakes language level accreditation test CertAcles (https://www.acles.es/index.php/en/). To-date, the research team has adapted the descriptors for L2 pragmatics and language complexity (Maíz-Arévalo & Méndez-García, in press; Díez-Bedmar, 2022b). The present study complements this work by zooming in on the construct of L2 accuracy.

For this paper, a sample of 100 FineDesc learner texts rated at the B1 level were exhaustively error annotated by the two authors following the latest version of the Louvain Error Tagging Manual (Granger, Swallow, Thewissen 2022) and in keeping with the methodology in Díez-Bedmar (2015, 2021). The corpus sample is made up of texts written in two different genres, namely email writing to a friend (50 texts, 6716 words) and creative story-telling (50 texts, 7812 words). Importantly, each text pair was written by the same learner. This enabled for novel analyses of two main types to be carried out: (1) the impact of task effects on L2 accuracy profiles was studied and showed that the frequency and type of errors found in informal emails and creative stories varied markedly. Task effects on L2 accuracy profiles have largely been ignored in learner corpus research to date, with the field presenting L2 accuracy levels as one homogeneous construct; (2) intra-learner variability was also captured: an objective accuracy-based ranking was produced for each learner and showed that the same learners do not necessarily perform with the same level of accuracy on both the email writing and the story telling tasks, thereby further highlighting the dynamicity of accuracy as a theoretical construct. This study also contributes to fine-tuning the descriptors of grammatical accuracy and vocabulary control at B1 level in the Companion Volume (Council of Europe, 2020) by making explicit references to the two text types, so that the descriptors are more nuanced and informative for Spanish L1 learners, teachers and testers.

References

Council of Europe (2001). *The common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe (2020). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume*. Strasbourg: Council of Europe Publishing.

Díez-Bedmar, M. B. (2015). The use of learner corpora to describe, teach and assess EFL writing: focus on the article system. In E. Castello, K. Ackerley & F. Coccetta (Eds.), *Studies in Learner Corpus Linguistics: Research and applications for Foreign Language Teaching and assessment* (pp. 37-69). Frankfurt: Peter Lang.

Díez-Bedmar, M. B. (2021). Error analysis. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 92-106). New York: Routledge.

Díez-Bedmar, M. B. (2022). Analysing learner language to inform the CEFR/CV descriptors: the FineDesc Project. Research into practice: towards a data-driven curriculum. One-day virtual (Zoom) conference - 23 March 2022.

Díez-Bedmar, M. B. (2022b). Informing the CEFR/CV Descriptors for Linguistic Range with Spanish L1 Learner Corpus Results: Focus on the Noun Phrase. International Congress on English Language Education and Applied Linguistics (ICELEAL 2022). Hong Kong, 6-9 December 2022.

Granger, S., Swallow, H., & Thewissen, J. (2022). *The Louvain Error Tagging Manual. Version 2.0*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain.

Maíz-Arévalo, C. & Méndez-García, C. (in press). “I would like to complain”: a study of the moves and strategies employed by Spanish EFL learners in formal complaint e-mails. *Intercultural Pragmatics*.

A Corpus-Based Analysis of 18th-Century American Grammars

Sophie Du Bois

The practice of writing grammars of English in the United States of America commenced much later than in its British ancestor. Considering a grammar to be American, when it was originally published in the United States, the first American grammar is Samuel Johnson's First Easy Rudiments of Grammar from 1765 (Lyman 1922, Nietz 1961, Alston 1965).

This study of 18th-century American grammars therefore aims to trace the origins of the practice of English grammar writing in the United States and the relations of American grammarians to their British ancestry. For this purpose, a corpus of five American grammars from the 18th century (ca. 87.000 tokens) was compiled.

Making use of the project-specific annotations in the corpus, onomastic references made to other persons are automatically extracted using a custom Python-based script. The resulting concordances are manually investigated and split into actual references and mentions within example sentences. Both categories are then further analyzed for the types of persons referenced using a system of person categories originally established for an analysis of 16th-century references in British grammars (Busse et al. 2021). The non-example references are further assigned categories initially developed for references in 19th-century British grammars (Busse et al. 2020). The results are displayed in a citation network (see White 2012), enabling a qualitative evaluation of references made. The observations are further enhanced by frequency analyses.

The investigations show that the majority of name-based references occur within the context of examples (59%), demonstrating the high value that is placed on demonstration by means of example sentences. Within these, religious and political figures are frequently mentioned, implying that not only religious texts (cf. Baron 1982, 126), but also political speeches were considered to be valuable moral guidance for the students.

Furthermore, although a sentiment of patriotism and independence passes through the country in the late 18th century, this sentiment has not yet reached early American grammarians in terms of their language ideologies. They predominantly reference British authors, implying a British dominance at the time in terms of authority on language matters, and confirming that the association of language and the American nation only culminated in the 19th century (Andresen 1990, 29-41).

References

Andresen, Julie Tetel. 1990. *Linguistics in America 1769-1924: A Critical History*. London: Routledge.

Alston, R. C. 1965. *English Grammar Written in English and English Grammar Written in Latin by Native Speakers: Vol. 1. A Bibliography of the English Language From the Invention of Printing to the Year 1800. A Systematic Record of Writings on English, based on the Collections of the Principal Libraries of the World.* Leeds: Arnold & Son.

Baron, Dennis E. 1982. *Grammar and Good Taste: Reforming the American Language*. New Haven, London: Yale University Press.

Busse, Beatrix, Ingo Kleiber, Nina Dumrukcic, 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. 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. Diskursmuster - Discourse Patterns 20. Berlin: De Gruyter.

Johnson, Samuel. 1765. *First Easy Rudiments of Grammar, Applied to the English Tongue*. New York: J. Holt.

Lyman, R. L. V. 1922. *English Grammar in American Schools Before 1850.* Chicago, Illinois: University of Chicago.

Nietz, John Alfred. 1961. *Old Textbooks: Spelling, Grammar, Reading, Arithmetic, Geography, American History, Civil Government, Physiology, Penmanship, Art, Music, as Taught in the Common Schools from Colonial Days to 1900*. Pittsburgh: American Book-Stratford Press, Inc.

Proficiency level influences EFL learners’ choice of genitive variant: Complementary evidence from corpus data and rating experiments

Tanguy Dubois

Whereas most previous research on alternation phenomena in learner language focus on the influence of the learners’ mother tongue on the choice of variant, this study investigates whether and how the use of the genitive variants, viz. the s-genitive (1) and the of-genitive (2), differs between native speakers and learners of English as a Foreign Language (EFL) at different proficiency levels.

(1) the city's name (TLC: 2_6_CH_32)

(2) the name of our city (TLC: 2_6_CH_53)

First, we collected many genitive observations from the Trinity Lancaster Corpus (Gablasova, Brezina & McEnery 2019, TLC), a three-million-word corpus consisting of recordings from an official language exam between a native speaker of British English and low-intermediate (B1) to advanced (C2) learners of English from several L1 backgrounds, including Chinese, Hindi, Russian, Spanish and Italian. The genitive observations (Ntotal = 2302) were annotated for various constraints, such as the length, animacy and definiteness of the constituents as well as the proficiency level of the speaker. The data was then analyzed via mixed-effects logistic regression, where the different constraints were allowed to interact with the speaker’s proficiency level. The results show that although native speakers and learners are similar, low-proficiency learners are less sensitive to possessor definiteness and possessor animacy, the latter of which is otherwise the strongest constraint of the genitive alternation. The corpus-based analysis was complemented with a scalar rating experiment, where participants had to rate the naturalness of either variant in different corpus excerpts (see Engel et al. 2022). The purpose of the experiment was to test whether the intuitions of native speakers and learners reflect the corpus predictions, and to corroborate the finding that low-proficiency learners are less sensitive to the animacy constraint. The results from the experiment, administered to 25 native speakers and 101 learners from seven different mother tongue backgrounds, show that the ratings of the participants correlate with corpus predictions, although this correlation is slightly weaker for low-proficiency learners. This result can partially be explained by the fact that low-proficiency learners are again less sensitive to the animacy constraint, which is a very strong predictor of genitive choice in the corpus model. We argue that learners are less sensitive to the animacy constraint because of a general preference for the of-genitive. At the same time, the constraint is difficult to learn because the animacy of the possessor noun phrase is largely determined by the context rather than formal cues.

References

Engel, Alexandra, Jason Grafmiller, Laura Rosseel & Benedikt Szmrecsanyi. 2022. Assessing the complexity of lectal competence: the register-specificity of the dative alternation after give. *Cognitive Linguistics* (aop).

Gablasova, Dana, Vaclav Brezina & Tony McEnery. 2019. The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research* 5(2). 126–158.

Exploring ways of distinguishing prize-winning novels from non-prize-winning ones

[Word in Progress report]

Jarle Ebeling

This work-in-progress report explores ways of linguistically distinguishing between prize-winning and non-prize-winning novels within the genre of general fiction. Previous research has shown how quantitative methods can be used to distinguish fiction from non-fiction, e.g., news and academic or learned writing (Stubbs & Barth 2003; Biber & □onrad 2012; Piper 2017). General fiction can also be distinguished from other literary genres with clear genre conventions, e.g., romance fiction, based on tokens (words and punctuation marks) alone, as shown in Figure 1, where romance novels, with file names ending in 'R', cluster together to the left in the diagram and general, prize-winning fiction novels, with file names ending in 'L', mostly cluster to the right. There are exceptions, e.g., the two romance novels C8S-R and MikGay1R, which cluster with general fiction and which would be interesting to look at more closely to find out why. Running the clustering algorithm on the general, non-prize-winning and the romance novels yields a picture similar to the one in Figure 1, but with more non-prize-winning novels clustering together with the romance novels.

A preliminary study, applying the same clustering algorithm to a corpus of general fiction only consisting of prize- and non-prize-winning novels (file names ending in 'L' vs. 'P') shows that the leaf nodes of the dendrogram tree do not fall neatly into two main clusters, as shown by Figure 2.

With the aim of investigating whether there are specific linguistic features that may better distinguish between prize-winning and non-prize-winning novels, this study applies methods developed within digital literary studies, e.g., most-frequent-words approaches (Hoover et al. 2014), and corpus linguistics, e.g., keyword approaches to phraseological patterning of texts (Bondi & Scott 2010) (see also Ashok et al. 2013, for similar approaches distinguishing successful – i.e., bestsellers – from less successful novels).

The main corpus used in the study consists of 50 novels, half of which have won the Booker prize while the other half has not. As a supplement, and to be able to compare the results of the main study with a different literary genre altogether, 25 romance novels have been chosen.

References

Ashok, Vikas G., Song Feng & Yejin Choi. 2013. Success with style: using writing style to predict the success of novels. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington: Association for Computational Linguistics, 1753–1764.

Biber, Douglas & Susan Conrad. 2009. *Register, Genre and Style*. Cambridge: CUP.

Bondi, Marina and Mike Scott. 2010. *Keyness in Texts*. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.41>

Hoover, David L., Jonathan Culpeper & Kieran O'Halloran. 2014. *Digital Literary Studies. Corpus Approaches to Poetry, Prose and Drama*. London: Routledge.

Piper, Andrew. 2017. Fictionality. *Journal of Cultural Analytics*, 2(2). <https://doi.org/10.22148/16.011>

Stubbs, Michael & Isabel Barth. 2003. Using recurrent phrases as text-type discriminators. A quantitative method and some findings. *Functions of Language* 10(1): 61–104.

Noun phrase modification in young learner writing

Kaja Evang

Greater phrasal complexity is seen as a trait of more mature academic writing, and the complexity typically increases as the academic level increases (Biber et al. 2011; 2014). However, more knowledge is needed about phrasal complexity at earlier stages in the linguistic development, as well as how it develops across L1 and L2 of the same writers.

This case study of texts from the 10th grade subset of the MULTIWRITE project corpus is conducted to inform a subsequent quantitative study on the Norwegian L1 and English L2 writing of 120 students in the final year of lower secondary school (10th grade). Pre- and postmodification in noun phrases are manually extracted from texts from regular school assignments in the subjects Norwegian and English. The texts are collected from four students from three different schools in order to avoid instruction bias (Norwegian students receive formal instruction in both Norwegian and English throughout the 10 years of obligatory education). The study aims to answer the following research questions:

1. How do Norwegian students modify noun phrases in their writing in their final year of lower secondary school?

2. Do their modifications differ in Norwegian (L1) versus English (L2)?

Previous research has shown that non-native university students underuse certain modification types when writing in English, compared to L1 students, and overuse evaluative modifiers (Tåqvist 2018; Larsson & Kaatari 2020). Although noun phrases have similar structures in English and Norwegian, a difference between the languages is found in postmodifying clauses, where finite relative clauses prevail in Norwegian, whereas non-finite alternatives more frequently appear in English (Elsness 2014). In a study of younger learners, 23% of Norwegian 7th grade students showed a similar syntactic complexity in English L2 as in Norwegian L1, while the majority produced less complex sentences in English (Drew 2001).

The result of the current study shows that there are differences between the students' writing in Norwegian and English, but that these are relative to type of modifier and to the students' individual writing styles. For example, one student uses adjectives only in English, and not in Norwegian, and another uses premodifying possessives and numerals only when writing in English. Not all students use infinitive postmodifiers, but those who do, do so both in English and Norwegian. A difference in complexity is not confirmed. The results illustrate the importance of distinguishing meticulously between modification types, and of considering each student's writing individually.

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. doi:10.5054/tq.2011.244483.

Biber, D., Gray, B., & Staples, S. (2014). Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. *Applied Linguistics*, *37*(5), 639–668.

Drew, I. (2001). A Comparison of Early Writing in Norwegian L1 and English L2. Paper presented at the *Annual Nordic Conference on Bilingualism* (8th, Stockholm, Sweden, November 1-3, 2001).

Elsness, J. (2014). Clausal modifiers in noun phrases: a comparison of English and Norwegian based on the Oslo Multilingual Corpus. S.O. Ebeling, A. Grønn, K. R. Hauge & D. Santos (eds.), *Corpus-based Studies in Contrastive Linguistics*. Oslo Studies in Language 6(1), 91-118.

Larsson, T. & H. Kaatari. (2020). Syntactic complexity across registers: Investigating (in)formality in second-language writing. *Journal of English for Academic Purposes, 45*, 100850.

MULTIWRITE project. https://www.hf.uio.no/ilos/english/research/projects/multiwrite/

Tåqvist, M. K. (2018). "A wise decision": Pre-modification of discourse-organising nouns in L2 writing. *Journal of Second Language Writing, 41*, 14–26. doi: 10.1016/j.jslw.2018.05.003.

The (re-)shaping of English discourse in video-mediated broadcast interviews with diplomats: a corpus-based analysis

Roberta Facchinetti

The language of interviews has been studied from a variety of perspectives and methodological approaches, particularly with reference to such fields as medicine (Mafi et al. 2018), journalism (Clayman and Heritage 2002), and politics (Piirainnen-Marsh 2005), bearing in mind race and culture (Tanaka 2004).

So far, however, little attention has been dedicated to the possible impact of video-mediated platforms on the shaping of questions and answers in broadcast interviews. Indeed, the recent dramatic shift to the online, which has affected the world of media possibly more than any other professional environment, calls for further research.

Bearing this in mind, the present study intends to delve into video-mediated broadcast interviews featuring adversarial questions and focusing particularly on diplomats as interviewees, a professional category that in turn has been understudied with reference to their specific discourse strategies. Two main research questions will be addressed:

(a) Are there any differences between video-mediated broadcast interviews and those carried out face to face?

(b) Bearing in mind that diplomats are trained to use their language favouring mediation and negotiation rather than assertiveness and aggressivity, how far do they respond to challenging questions in video-mediated vs face-to-face interviews?

To carry out the study, I will analyze a subset of the "InterDiplo Corpus", which covers broadcast interviews of professional journalists to diplomats, politicians and international experts from different cultural backgrounds and is specifically tagged to concentrate on the question-answer interface between interviewer and interviewee. The subcorpus taken into consideration covers two equal sets of face-to-face and video-mediated broadcasts between journalists and diplomats to compare non-linguistic and linguistics aspects of their dialogue and focuses on different types of questions and answers (open, close, choice and requests).

The data yielded by the analysis point to a set of differences (as well as some similarities) between video-mediated and face-to-face interviews both in non-verbal communication (gestures, facial expressions, eye/hand movements), and in verbal discourse, particularly with reference to the use of ‘polarized lexicon’ expressing adversarial stance (ex: right/wrong), to ‘loaded lexicon’ (ex: aggression/attack vs military action), as well as to the rhetorical strategies of evasiveness and presupposition. Such differences pertain both to interviewers and to interviewees in different degrees, testifying to increased verbal aggressiveness in this type of video-mediated dialogue. Overall the data point to video-mediated broadcast interviews as an evolving type in discourse, featuring partly unexpected traits particularly when it comes to the language of diplomats.

References

Clayman Steven and John Heritage (2002) *The News Interview: Journalists and Public Figures on the Air.* Cambridge: Cambridge University Press.

Mafi, John, Macda Gerard, Hannah Chimowitz, Melissa Anselmo, Tom Delbanco, Jan Walker (2018) “Patients Contributing to Their Doctors’ Notes: Insights From Expert Interviews.” *Annals of internal medicine* 168/4: 302–305.

Piirainnen-Marsh, Arja (2005) “Managing Adversarial Questions in Broadcast Interviews”. *Journal of Politeness Research*. 1/2: 193-217.

Tanaka, Lidia (2004) *Gender, Language and Culture: A Study of Japanese Television Interview Discourse*. Amsterdam Philadelphia: John Benjamins.

Syntactic Complexity Development of Intermediate L2 English: A longitudinal, corpus-based study

Sandra Götz and Philine Tschirner

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 & □rossley 2015) have led to very detailed descriptions of L2 English complexity development (e.g. Lu 2010; Biber et al. 2011; Kyle & □rossley 2015; Kyle, □rossley & 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 generalizations are often hard to make. Additionally, despite the comparatively long research tradition of complexity research, truly longitudinal corpus-based studies tracing the complexity development of large learner groups by taking into consideration the effect of learning context variables remain very rare (cf., however, Kyle, Crossley & Verspoor 2021, who albeit 'only' analyze 9 learners over two academic years).

Against this background, in the proposed work-in-progress report, we would like to introduce a new research project that investigates if/how syntactic complexity develops in a large group of intermediate German learners of English over four school years. More specifically, the proposed project 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?

In order to answer these research questions, we will analyze 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. The corpus data are 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 values will then be manually checked for their accuracy and potentially be corrected (cf. Châu & Bulté 2022). The dataset will then be complemented with the metadata from MILE's learner profiles (e.g. gender, age or grade), before it will be subjected to a statistical data analysis using mixed effects regression modelling (e.g. Gries 2015) with the software package R (R core team 2022), so that we will be able to control for potential effects of individual learner variation and learning context variables on complexity development. 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; cf. also Larsen-Freeman 1978), whereas other variables seem to have varying effects that will need closer inspection.

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. & B. Bulté (2022). A comparison of automated and manual analyses of syntactic complexity in L2 English writing. International Journal of Corpus Linguistics, available online https://www.jbe-platform.com/content/journals/10.1075/ijcl.20181.cha

Gries, Stefan Th. (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, Rolf. 2015. The Marburg Corpus of Intermediate Learner English (MILE). In Learner Corpora in Language Testing and Assessment, Marcus Callies & Sandra Götz, eds. Amsterdam: John Benjamins. 13-34.

Kyle, Kristopher & Crossley, Scott A. 2015. Automatically assessing lexical sophistication: Indices, tools, findings & application. TESOL Quarterly 49(4): 757–786.

Kyle Kristopher & Crossley, Scott A. & Marjolijn Verspoor. 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, Diane. 2009. Adjusting expectations: The study of complexity, accuracy, and fluency in Second Language Acquisition. Applied Linguistics 30(4): 579–589.

R Core Team. (2022). A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/.

Overhauling keyness analyses with three dimensions & word2vec: contrasting three Asian varieties of English

Stefan Gries

One central corpus-linguistic methods is keyness analysis, the identification and interpretation of words of a target corpus (*T*) that, compared to their occurrence in a reference corpus (*R*), are key/characteristic for *T*. Statistically, this is most often done with log-likelihood ratio tests comparing the token frequencies of each word type in *T* and *R*. In terms of application practice, *T* and *R* usually differ regarding their topic areas (e.g. to identify expressions relevant for topic areas in language teaching contexts) or registers/genres (e.g. to identify expressions typical of certain communicative contexts/situations). However, ever since, e.g., Leech & Fallon (1992), studies have also used keywords analysis to compare corpora representative for different cultures; Leech & Fallon (1992) themselves compared British and American written English based on the LOB and Brown corpora, but more recent work has also targeted varieties of English (e.g. Mukherjee & Bernaisch 2015 or Collins 2021).

We will outline a set of suggestions for how to conduct better keyness analyses and exemplify it on the basis of a cultural keyness analysis of newspaper corpora for Indian, Pakistani, and Sri Lankan English. We will first distinguish two keywords scenarios: (i) the prototypical one, which seeks to identify keywords in a bottom-up fashion and (ii) a less typical one in which a researcher is interested in a top-down determined set of words and their differences across corpora. As for (i), we discuss a new way to compute keywords in general, which is an extension of Gries (2021) and considers word frequency but also a frequency- and a dispersion-based component of keyness; we will discuss the results of comparing each variety

to the other two in the SAVE 2020 corpus.

As for (ii), we exemplify how one can use distributional-semantics kinds of approaches to either disambiguate keywords yielded by the first approach or to determine how words used in each variety are associated with different kinds of ‘collocates’ (as returned by a collocational interpretation of cosine similarity) and what that indicates. Replicating Mukherjee & Bernaisch (2015), we show how the IndE, PakE, and SriE newspaper corpora differ substantially in how they discuss *terror*:

- in IndE, there is an overrepresentation of military terms such as *artillery*, *combating*, *commando*, *drone*, *enemy*, *grenade*, *occupier*, *squads*, and *submarine*; the IndE news coverage of *terror* might be concerned a lot with how terror is performed and what (military) reactions it prompts;
- in PakE, there is an overrepresentation of terms that involve misleading communication such as *indoctrination*, *intimidation*, *misinformation*, *propaganda*, *provocation*, *ruse*, *spout*, and *stereotyping*;
- in SriE, there is an overrepresentation of general religious terms (e.g., *churches*, *cult*, *fanatic*, *muslim*, , *shiite*, *sunni*, *taslim*, *religiosity*, *salafi*, *salafist*, and *extremists*) but also the word family of *jihad*; the SriE news coverage considers (extreme) religiousness as relevant to terror.

References

Collins, P. 2021. Cultural keywords in World Englishes: A GloWbE-based study. *ICAME Journal*, 45, 5–35.

Gries, St.Th. 2021. A new approach to (key) keywords analysis: using frequency, and now also dispersion. *Research in Corpus Linguistics* 9/2, 1–33.

Leech, G. & Fallon, R. 1992. Computer corpora – What do they tell us about culture? *ICAME Journal*, 16, 29–50.

Mukherjee, J. & Bernaisch, T. 2015. Cultural keywords in context: A pilot study of linguistic acculturation in South Asian Englishes. In Collins, P. (ed.), *Grammatical Change in English World-Wide*, 411–35. Amsterdam: Benjamins.

Against level-3 frequency comparisons: why and what to do instead

Stefan Gries

In the last few decades, much work in corpus linguistics has attempted to discover, and then interpret, differences in the frequencies of use of linguistic elements such as certain words, patterns, constructions, discourse features, etc. It is probably fair to say that such studies were particularly frequent in (i) learner corpus research, (ii) corpus-based varieties research, and (iii) sociolinguistically motivated studies. For instance, as for (i), many studies have tested and discussed the differences in how often certain words or constructions are used in corpus data from native speakers vs. corpus data from learner from different L1 backgrounds (e.g., Chen 2013 on phrasal verb use by native speaker and Chinese learners); as for (ii), many studies have tested and discussed how often certain words or constructions are used in corpus data representing different inner- and outer-circle varieties (e.g. Bruckmaier 2017 on GET in World Englishes); as for (iii), studies have explored how often certain words or constructions are used by men and women in corpora (e.g., Argamon et al. 2003).

This paper will make the admittedly bold claim that any such study can in fact *by definition* unable to 'prove' what is often their main points, namely that the distributional differences found are in fact due to the one hypothesized explanatory variable(s) of *L1*, *VARIETY*, and *GENDER*; note that this is true even if the distributional differences are very or even highly significant (as per the usual chi-squared or log-likelihood tests) and come with a decent effect size (e.g., φ or Cramer's V). To substantiate this claim, I will first discuss some simple terminology from the family of methods known as multi-level modeling, namely the distinction between level-1, level-2, … level-*n* variables and how this set of levels maps onto the vast majority of corpus-linguistic studies: Specifically, *L1* and *VARIETY* will be shown to correspond to level-3 variables whereas *GENDER* corresponds to a level-2 variable. Second, I will then demonstrate using both made-up and authentic data (from two case studies, one from learner corpus research, one from research on South Asian varieties of English) how studies using only the above kinds of variables cannot distinguish the effect of their favored predictors from the effect of local/contextual level-1 variables. Third, I will exemplify (with just one case study) how claims regarding differences due to *L1*, *VARIETY*, and *GENDER* need to be explored quantitatively instead.

References

Argamon, S.; Koppel, M.; Fine, J.; Shimoni, A. R. 2003. Gender, Genre, and Writing Style in Formal Written Texts. *Text & Talk*, 23/3, 321–46.

Bruckmaier, E. 2017. *Getting at GET in World Englishes*. Berlin/Boston: de Gruyter.

Chen, M. 2013. Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics*, 18/3, 418–42.

A Consistent Approach to Annotating and Modelling Metadiscourse for Non-coders

Wenwen Guan

Although labelling techniques of linguistic forms have been extensively applied, such as POS tagging and syntactic parsing, the annotation of metadiscourse still heavily relies on manual efforts. The nature of metadiscourse hampers automated annotation. Firstly, metadiscourse has various forms ranging from one word to a whole sentence. Secondly, it is multifunctional, so annotators need to identify its function according to the context. Traditionally, there are two methods to annotate metadiscourse. The “thin” method retrieves metadiscourse candidates on a pre-defined list concluded by Hyland (2005). The process can be highly automated; however, it may ignore possible metadiscourse that is not on the list and may not produce context-sensitive results. The other method is called the “thick” approach, which requires annotators to go through each metadiscourse candidate in the context (Ädel & Mauranen, 2010). Despite satisfactory annotation accuracy, the thick approach is costly and time-consuming.

Faced with the dilemma, we investigated an AI-assisted approach to annotating metadiscourse. The approach is performed using Prodi.gy, an all-in-one annotation tool which has embedded machine learning functions. The present study demonstrates the intact workflow of our metadiscourse annotation project from data preparation to result analysis. This is an advantage over conventional corpus tools and NLP methods because all steps can be realized on one platform without any prior knowledge of coding.

What genuinely makes Prodi.gy distinctive is its suggester function. In other words, Prodi.gy users can train a model with a small amount of manually annotated data, and it will suggest labels for the rest of the data. To examine how much the function contributes to the annotation process, we observed the accuracy of suggestions and the time saved by suggestion-assisted annotation with different percentages of manually labelled data. Moreover, we integrated Hyland’s list as “patterns to be matched” into the suggester function, hence showing the candidates at one go without retrieving them one by one. Besides, we conducted a qualitative analysis of some tricky metadiscourse candidates and gave feedback to annotators in time.

Prodi.gy turns out to be a nice trade-off between fully manual annotation and automatic annotation. The suggestions significantly shorten annotation duration and result in sufficient accuracy. Timely feedback also improves annotators’ experience. Our workflow is replicable for studies on other functional linguistic subjects like metadiscourse, and also has the potential to be applied to varied types of linguistic data, such as audio data and sign language data.

Developing the LC22 learner corpus of academic English writing: an exploration of challenges and solutions

Sharon Hartle

Developing a learner corpus involves a series of challenges (McEnery et al., 2019), such as choice of reference corpus (ns or nns), involving the careful assessment of L2 language use, but, when using an L1 as a reference norm, also analysis of L1 usage (Gablasova et al., 2017). The labour-intensive nature of the compilation and annotation processes when adopting a computer aided analysis (CEA) approach may also be an issue, together with a potentially reduced effectiveness of statistical analysis as learner corpora may be smaller in size. This presentation briefly describes the development of the Learner Corpus 22 (LC22). It was developed on Sketch Engine, as part of a pilot study, primarily, to determine the local EAP needs at the University of Verona, and to inform the learning design of Academic Writing courses, a subject of wide debate (Flowerdew & Peacock, 2001; Hyland, 1998; Tang, 2012). LC22 consists of two sub-corpora, one of summary writing and the other of discursive discussion writing, based on the production of post-graduate students with the aim of answering two main questions:

- What are the principal strengths and weaknesses in advanced learner academic writing?
- What resources may be developed to aid lexical acquisition and competence?

The main part of the presentation, however, explores the compilation and particularly the annotation process outlining some of these challenges (which norms to use, how to reference them) and how solutions were found. Although some of these problems are specific to our local context, such as the relative difficulty levels of Latin-based lexis, they have a wider application insofar as every local context may meet similar challenges. LC22 was created firstly by means of XML coding, which enabled tagging for error and successful usage at the same time, and then analyzed in Text Inspector and the Sketch Engine.

References

Flowerdew, J., & Peacock, M. 2001. Issues in EAP: A preliminary perspective. In J. Flowerdew and M. Peacock (eds.), *Flowerdew and Peacock Research Perspectives on English for Academic Purposes.* Cambridge: Cambridge University Press, 8-24.

Gablasova, D., Brezina, V., & McEnery, T. 2017. Exploring Learner Language Through

Corpora: Comparing and Interpreting Corpus Frequency Information. *Language Learning* 67:S1, 130-154.

Hyland, K. 1998. *Hedging in Scientific Research Articles.* Amsterdam and Philadelphia: John Benjamins Publishing Company.

McEnery, T., Brezina, V, Gablasova D., & Banerjee J. 2019. Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics* (39) pp. 74-92.

Tang, R. 2012. The Issues and Challenges Facing ESL/EFL Academic Writers in Higher Education Contexts: An Overview. In *Academic Writing in a Second or Foreign Language. Issues and Challenges Facing ESL/EFL Academic Writers in Higher Education Contexts.* London: Continuum Publishing, 1-20.

Text Inspector. https://textinspector.com/ last accessed 21/03/2023

The Sketch Engine. https://www.sketchengine.eu/ last accessed 21/03/2023

Plural Reference in Non-Canonical English Second Person Pronouns

David Hernández-Coalla

Ever since *you* encroached on the remaining functions of *thou* in Early Modern English, it has been widely accepted that there is a single second person pronoun in standard English for both singular and plural reference. Although English speakers feel no need to add another personal pronoun to the paradigm (De Vogelaer 2007) in order to make a number distinction, such a categorical conclusion may not be backed by empirical data. A quick search on the GloWbE (Davies 2013) reveals that in certain geographical varieties *you* is sometimes used with a verb, seemingly inflected for the singular with the traditional third person singular ending. In addition, the OED lists several non-canonical pronominal forms with a second person reference: *youse*, *yiz*, *oonu*, etc. Using the GloWbE, the OED and a list of descriptive grammars, the aim of this paper is to determine whether these non-standard forms are actually used in different varieties of English and whether they have a plural reference in contrast to singular *you*. Special attention will be given to periphrastic pronominal forms such as *you all*, *y'all*, *you guys* and *you lot*, since these appear to be preferred over other one-word pronominal alternatives. In particular, two questions will be addressed regarding these forms. First, it is imperative to clarify if they qualify in fact as personal pronouns or they are periphrastic combinations of standard *you* and a quantifier or a noun (Valentínyová 2015). To this purpose, this paper will assess their chances of having undergone semantic bleaching and their contexts of use. Second, *you all* will be analysed in depth in all the varieties in which it is attested. Despite being typically described as an American form especially common in the Southern states of the country, its high frequency in many national varieties begs the question of whether its uses are similar to those found in American English, which would point to its adoption from this variety. These questions will be answered with the data retrieved from this work in progress.

References

Davies, Mark. (2013). *Global Web-Based English Corpus (GloWbE)*. Available at: https://english-corpora.org/glowbe/

De Vogelaer, Gunther. (2007). Innovative 2pl.-pronouns in English and Dutch. “Darwinian” or “Lamarckian” change? *Papers of the LSB, 15.* Belgische Kring Voor Linguïstiek. https://sites.uclouvain.be/bkl-cbl/wp-content/uploads/2014/08/vog2007.pdf

Oxford English Dictionary. (December 2022). *OED Online*. Oxford University Press. https://oed.com

Valentínyová, Kristína. (2015). Non-standard forms of the pronoun “you” in English. Prague: Univerzita Karlova [Unpublished Bachelor’s Dissertation]. https://dspace.cuni.cz/bitstream/handle/20.500.11956/66219/BPTX-2013_2_11210_0_345119_0_154697.pdf?sequence=1&isAllowed=y

Young learners' use and development of linking devices [Work-in-progress report]

Stine Hulleberg Johansen

Linking words and phrases are devices which indicate a relationship between units of discourse (Biber et al., 1999) and constitute a language feature which has proven difficult for learners to master. Although linking devices in L2 English have been extensively studied (see e.g., Granger & Tyson, 1996; Bolton et al., 2002; Köroğlu, 2019), the vast majority of studies have focused on advanced learners in tertiary education. There is limited knowledge about the earlier stages of learner writing, and on (individual) development over time. Furthermore, many studies have focused on over-/underuse and compared learner corpora to more general reference corpora of L1 English (e.g., Milton & Tsang, 1993). There are comparatively fewer studies of misuse, and studies where L2 corpora are compared to corpora of L1 learners/novice writers.

The present study aims to address these gaps by investigating linking devices used by Norwegian learners of English in secondary education over one year. The study investigates what type of linking devices these learners use, whether they signal addition, contrast, concession, etc., and how their use develops. Ultimately, the results will be qualitatively compared with those from an ongoing study by Philip Durrant at Exeter University of L1 learners, which uses data from the Growth in Grammar Corpus (Durrant & Brenchley, 2018). Comparing L2 with L1 learners makes it possible to distinguish features of L2 writing from general writing development.

The study is a part of a larger research project investigating the writing development of Norwegian students in a corpus under construction consisting of written texts in the students' first (Norwegian), second (English) and third (German, French or Spanish) language during their first year of upper secondary school. This study utilizes a sub-corpus consisting of 192 English texts (101,572 word tokens) produced by 31 students, with an average of 6 texts from each student. A list of 267 unique linking expressions from the linking literature (Halliday & Hasan, 1976; Quirk et al., 1985; Biber et al., 1999; Carter & McCarthy, 2006; Liu, 2008; Larsen-Freeman & Celce-Murcia, 2016; Yin, 2016) will serve as the basis for the search (Durrant, pers. comm.) Mixed-effects regression modelling will be used to illustrate the development of the use of different types of linking devices in individual students and for the group as a whole, before the qualitative comparison with L1 learners.

The study hopes to provide more detailed knowledge about the development of linking devices in learners' writing.

References

Biber, D., Conrad, S. and Leech, G. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Bolton, K., Nelson, G., & Hung, J. (2002). A corpus-based study of connectors in student writing. *International Journal of Corpus Linguistics*, 7(2), 165-182.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English. Cambridge*; New York: Cambridge University Press.

Durrant, P. & Brenchley, M. (2018). Growth in Grammar Corpus. Retrieved from gigcorpus.com. (registration require – contact Philip Durrant for access details: p.l.durrant@exeter.ac.uk)

Durrant, P. personal communication, January 9, 2023.

Granger, S. & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speak-ers of English. *World Englishes*, 15 (1), 17-27.

Halliday, M. A. K. & R. Hasan. (1976). *Cohesion in English*. Longman.

Köroğlu, Z. (2019). A corpus-based analysis: The types of transition markers in the MA theses of native speakers of English and Turkish speakers of English. *Journal of Language and Linguistic Studies*, 15(2), 496-507. Doi: 10.17263/jlls.586157

Milton, J., & Tsang, E. S. C. (1993). A corpus-based study of logical connectors in EFL students' writing: directions for future research. In R. Pemberton & E. S. C. Tsang (Eds.), *Studies in lexis* (pp. 215-246). Hong Kong: The Hong Kong Unversity of Science and Technology Language Centre.

Larsen-Freeman, D. & M. Celce-Murcia (2016). *The Grammar Book: Form, meaning, and use for English language teachers*. Boston, National Geographic Learning.

Liu, D. (2008). "Linking adverbials: An across-register corpus study and its implications." *International Journal of Corpus Linguistics* 13(4): 491-518.

Quirk. R; Greenbaum, S.; Leech, G. & Svartvic, J. (1972*). A Grammar of Contemporary English*. London: Longman.

Yin, Z. (2016). "Register-specific meaning categorization of linking adverbials in English." *Journal of English for Academic Purposes* 22: 1-18.

***It would take a lot for them to take to the streets*– The introductory-*it* pattern in the Newspaper Language of South Asian Varieties of English**

Kathrin Kircili and Sandra Götz

Both spoken and written discourse underly a versatility of (information) structural principles easing both the production and perception of a given message which, if bearing the *canonical* constituent order (i.e. Sn-V-X), would be much less felicitous. These include, e.g., the *end-weight principle* (Behaghel 1909), the *end-focus principle* (Huddleston & Pullum 2002: 1371) or that of discourse familiarity (□lark & Haviland 1977; Prince 1981). One structure that conforms to each of these concepts is the so-called introductory-*it* pattern (also known as (*it*-)extrapositioning (Kaltenböck 2004) or anticipatory-*it* (Biber et al. 1999; Hewings & Hewings 2002) in case of which lengthy notional subjects or objects are moved towards the end of a construction while their grammatical position is filled by an expletive *it*:

1) It[Sg] is[V] nice[C] to see you.[Sn]

2) He[S] must find [V]it[Og] boring[C] to listen to her.[On]

Since its first discussion by Jespersen (1927), it has received quite a bit of scholarly attention in ENL and EFL research. Findings in the former domain include its employment to achieve particular semantic goals, including the expression of necessity, importance, ease or difficulty (Biber et al. 1999). From a structural perspective, *that*-clauses are most commonly extraposed (ibid; Zhang 2015) while *ing*-clauses are generally preferred in spoken discourse (Quirk et al. 1985: 1393). In EFL, the pattern is generally overrepresented (□allies 2009; Larsson 2017), with varying structural as well as functional distributions and preferences (ibid; Hewings & Hewings 2002), such as to "[comment] on, [evaluate], or [hedge] the following clause" (ibid: 372).

In ESL, non-canonical structures in general have been described as common characteristics of 'New Englishes' (Platt et al. 1984; Sharma 2012); however, although there are studies on a number of these patterns (Lange 2012; Winkle 2015; Leuckert 2019), to the best of our knowledge, no in-depth comparative account of the structural and functional peculiarities of introductory-*it* in the whole South Asian region has been attempted. We therefore intend to answer the following research questions:

1) Are there differences between the SAVEs (compared to BrE) in using the introductory-*it* pattern regarding a. frequencies b. structural preferences c. functional preferences

2) Which factors influence its choice (as opposed to the canonical counterpart)?

Making use of the SAVE corpus (Bernaisch et al. 2011) and the BNC news section, approximately 56,800 sentences were manually annotated for a number of variables (e.g. SENTENCE_LENGTH, STRUCTURE_TYPE, SEMANTIC_TYPE) before submitting them to regression modeling. The results reveal both similarities –such as the increased likelihood of the structure to occur with an increasing sentence length– and clear differences in frequency and structural as well as semantic preference, e.g. the fact that Nepali English clearly favors *to*-infinitives over *that*-clauses or the fact that, given our focus on newspaper language, the verbal and post-verbal (and thus pre-clausal) constituents clearly indicate the intention to provide neutral and strictly informative accounts. The results will be discussed against the background of previous findings on the topic.

References

Behaghel, O. (1909): “Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern“, *Indogermanische Forschungen* 25, 110-142.

Bernaisch, T., C. Koch, J. Mukherjee and M. Schilk (2011). Manual for the South Asian Varieties of English (SAVE) Corpus: Compilation, Cleanup Process, and Details on the Individual Components. Giessen: Justus Liebig University, Department of English.

Biber, D., S. Johannson, G. Leech, S. Conrad & E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Callies, M. (2009). *Information Highlighting in Advanced Learner English: The syntax-pragmatics interface in second language acquisition*. Amsterdam/Philadelphia: John Benjamins Publishing.

Clark, H. H. & S. E. Haviland (1977). “Comprehension and the Given-New Contract”, *Discourse Production and Comprehension*, eds. R.O. Freedle. Norwood: Ablex Publishing Corporation, 1–40.

Hewings, M. & A. Hewings (2002). ““It is interesting to note that…”: A comparative study of anticipatory ‘it’ in student and published writing”, *English for Specific Purposes*, 21 (4), 367–383.

Huddleston, R. & G. K. Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Jespersen, O. (1927). A modern English grammar on historical principles. Vol. III. Heidelberg: Carl Winter.

Kaltenböck, G. (2004). *It-extraposition and non-extraposition in English: a study of syntax in spoken and written texts*. Wien: Braumüller.

Lange, C. (2012). *The Syntax of Spoken Indian English*. Amsterdam: John Benjamins.

Larsson, T. (2017). “A functional classification of the introductory it pattern: Investigating academic writing by non-native-speaker and native-speaker students.” *English for Specific Purposes* 48: 57-70 .

Leuckert, S.(2019). *Topicalization in Asian Englishes: Forms, Functions, and Frequencies of a Fronting Construction*. London: Routledge.

Platt, J. T., H. Weber & M. L. Ho. (1984). *The New Englishes*. London: Routledge.

Prince, E. (1981). “Toward a Taxonomy of Given-New Information”, *Radical Pragmatics*, ed. P. Cole. New York: Academic Press. 223-255.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. Harlow: Longman.

Sharma, D. (2012). “Shared Features in New Englishes.” R. Hickey, ed. Areal Features of the Anglophone World. Berlin: De Gruyter, 211-232.

Winkle, C. (2015). *Non-canonical structures, they use them differently: Information packaging in spoken varieties of English*. Dissertation, Albert–Ludwigs-Universität Freiburg.

Zhang, G. (2015). "It is suggested that…or it is better to…? Forms and meanings of subject it-extraposition in academic and popular writing." *Journal of English for Academic Purposes* 20, 1-13.

Exploring features of elaborate vs. compressed language in spoken academic discourse: An SEM approach

Maria Kostromitina, Larissa Goulart and Tove Larsson

Multiple studies of register variation have reported that spoken discourse is characterized by features of elaboration, or clausal features, while written discourse is characterized by features of compression, or phrasal features (e.g., Biber, 2006; Biber et al. 2022). However, there is recent evidence calling for more specificity, as written domains, such as different kinds of academic writing, have been found to vary considerably in this respect, depending on which disciplines and registers are studied (Gardner et al., 2019; Goulart, 2022). For example, writing in the humanities and writing that demonstrates content knowledge exhibit frequent use of features of elaborate discourse (Goulart, 2021). This paper tests whether such mode-internal variation extends to spoken registers as well, and seeks to answer the following question: To what extent do register and discipline affect features of clausal elaboration vs. phrasal compression in spoken academic discourse?

We use the Biber Tagger (Biber, 2006) to extract features commonly associated with compression (N+N, ADJ+N, N+PP, and of-genitives) and clausal elaboration (verb complement clauses, noun complement clauses, and adverbial clauses) in 50 classroom encounters from four disciplinary groups (arts and humanities, social sciences, life science, and physical sciences) from the British Academic Spoken Corpora (BASE). Following Biber et al. (2020), we consider registers a continuum and operationalize it using the level of interactivity of the classroom encounter. Based on previous research on writing, we hypothesize that

(1) Both register and discipline will play a role in the linguistic variation of classroom encounters

(2) Higher interactivity will lead to more features of elaboration; lower interactivity will lead to more features of compression

(3) The humanities will have more features of elaboration; the physical sciences will have more features of compression

Using measured variable path analysis from the structural equation modeling framework (Larsson et al., 2021), we fitted competing models to test our hypotheses. The results support our hypothesis that both register and discipline play a role (CFI: 0.981, RMSEA: 0.104, SRMR: 0.037), and that monologic contexts contained more features of compression. The results for discipline did not fully meet our expectations in that disciplines in the sciences used more premodifying nouns and subordinate clauses, whereas the humanities exhibited more attributive adjectives, prepositional phrases, and verb complement clauses. The results support a need for a more nuanced view of the role of elaborate vs. compressed features in spoken discourse, one that also considers text-external factors.

References

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* Amsterdam: Benjamins.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2022). *The Register-Functional approach to grammatical complexity: Theoretical foundation, descriptive research findings, applications*. Routledge.

Biber, D., Egbert, J., & Keller, D. (2020). Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16(3).

Gardner, S., Nesi, H., & Biber, D. (2019). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*, *40*(4), 646–674.

Goulart, L. (2021). Register variation in L1 and L2 student writing: A multidimensional analysis. *Register Studies, 3*(1), 115–143.

Goulart, L., Biber, D., & Reppen, R. (2022). In this essay, I will…: Examining variation of communicative purpose in student written genres. *Journal of English for Academic Purposes, 59*, 101159.

Larsson, T., Plonsky, L., & Hancock, G. R. (2021). On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory*, *17*(3), 683–714.

Negative polarity items in Inner and Outer Circle Englishes

Claudia Lange

One categorical difference between Inner and Outer Circle Englishes seems to be the absence of nonstandard forms of negation: double negation or 'ain't' for 'have/be' are conspicuously absent from L2 varieties of English (Anderwald 2012). This paper will extend the scope of research on negation across varieties of English by focussing on the forms, frequencies and distribution of "negatively-oriented polarity-sensitive items (NPIs)" (Huddleston & Pullum 2002: 823). These include, among others, "the any-class of items", e.g. 'any more, anybody' (Huddleston & Pullum 2002: 823), as well as "minimizers" and "verbal idiomatic expressions" (Tovena 2020: 392) of the form 'neg-V a N', such as 'I did not drink a drop/have a clue'. Since especially the latter are more likely to occur in spoken language, data from recent corpora of spoken language are considered, namely the spoken BNC 2014 (Love et al. 2017) as representative of an L1 variety, and the direct conversation files of the roughly contemporaneous ICE-Sri Lanka and ICE-Nigeria, representing two L2 varieties embedded in different contact scenarios and sociocultural settings. Preliminary data analysis shows that the NPI class of minimizers/idiomatic expressions, e.g. 'don't give a shit/fuck' in present-day spoken British English, is extremely rare in the two L2 Englishes under scrutiny, raising the question whether L2 varieties opt for different realizations of L1 idiomatic expressions for emphatic negation.

References

Anderwald, Lieselotte (2012), "Negation in Varieties of English." In: Hickey, Raymond (ed.), *Areal Features of the Anglophone World*. Berlin: de Gruyter Mouton, 299-328.

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

Hoeksema, Jack (2010), "Negative and Positive Polarity Items: An Investigation of the Interplay of Lexical Meaning and Global Conditions on Expression." In Horn, Laurence R. (ed.), *The Expression of Negation*. Berlin: de Gruyter Mouton, 187-224.

Horn, Laurence R. (1989), *A Natural History of Negation*. Stanford: CSLI Publications.

Huddleston, Rodney D. and Pullum, Geoffrey K. (2002), "Negation." In Huddleston, Rodney D. and Pullum, Geoffrey K. (eds.), *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press, 785-850.

Love, Robby, Dembry, Claire, Hardie, Andrew, Brezina, Vaclav and McEnery, Tony (2017), "The Spoken BNC2014: designing and building a spoken corpus of everyday conversations." *International Journal of Corpus Linguistics* 22(3): 319-344. DOI: 10.1075/ijcl.22.3.02lov

Meyler, Michael (2012), "Sri Lankan English." In Kortmann, Bernd & Lunkenheimer, Kerstin (eds.), *The Mouton World Atlas of Variation in English.* Berlin: de Gruyter Mouton, 540-547.

Miestamo, Matti (2017), "Negation." In Aikhenvald, Alexandra Y. and R. M. W. Dixon (eds.), *The Cambridge Handbook of Linguistic Typology*. Cambridge: Cambridge University Press, 405-439.

Taiwo, Rotimi (2012), “Nigerian English.” In Kortmann, Bernd & Lunkenheimer, Kerstin (eds.), *The Mouton World Atlas of Variation in English*. Berlin: de Gruyter Mouton. 410-416.

Tovena, Lucia M. (2020), “Negative Polarity Items.” In Déprez, Viviane & Espinal, M. Teresa (eds.), *The Oxford Handbook of Negation*. Oxford: Oxford University Press, 391-406.

A conceptual replication of the Multi-Dimensional Model of General Spoken and Written English (Biber 1988): Challenges, limitations, and potential solutions

Elen Le Foll

The term 'reproducibility' – typically defined as the ability to obtain the same results as an original study using the authors' data and code – is often distinguished from 'replicability', which refers to the obtention of compatible results following the same method but with different data (Berez-Kroeker et al. 2018). In this paper, we argue that if Biber's (1988) multidimensional (MD) model of General Spoken and Written English is to be generalisable and used as a baseline model in additive MD analyses (Berber Sardinha et al. 2019), its dimensions ought to be replicable with a new corpus – provided that this corpus was compiled to be representative of "General English", too. Moreover, if the 1988 model's dimensions represent underlying functional dimensions of variation in "General English", these ought to be replicable even with a different set of linguistic features and/or different feature operationalisations. Hence, the present study aims to replicate the 1988 model by conducting a full MD analysis of "General English" using a different corpus and tagset with a view of assessing the original model's validity, stability, and reliability.

Whilst Biber's (1988) MD model is based on counts of 67 lexico-grammatical and semantic features as identified by the 1988 version of the Biber Tagger, this replication relies on 75 linguistic features identified by the MFTE (Perl version; Le Foll 2021) in a stratified sample of the BNC2014 (Brezina, Hawtin & McEnery 2021; Love et al. 2017). We compare the feature loadings of the five-factor solution generated from the BNC2014 Baby+ data (see Table 1) to Biber's (1988) six dimensions and their feature loadings. Dimension scores for each text are also computed and plotted (see Fig. 1). The results are then compared to the dimension scores of the different (sub)registers of the original corpus.

Broadly speaking, this new MD model based on the BNC2014 Baby+ successfully replicates Biber's (1988) Dimensions 1, 2, 4 and 5. Some of the differences observed in the composition of the dimensions can be traced back to the different normalisation baselines. Additionally, a new dimension emerged, which captures a type of register variation not found in Biber's (1988) model because it largely serves to distinguish a text type that was absent from Biber's (1988) corpus. This points to the inherent difficulty of compiling a corpus that accurately captures the full breadth of register variation in "General English".

Based on these mixed results, we highlight challenges in producing robust results using Biber's (1988) MD framework that arise at all stages of the complex MD analysis framework. We propose various solutions to overcome these and advocate for greater transparency in the parameters used in future MD analyses.

References

Berber Sardinha, Tony, Marcia Veirano Pinto, Cristina Mayer, Maria Carolina Zuppardi & Carlos Henrique Kauffmann. 2019. Adding Registers to a Previous Multi-Dimensional Analysis. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 165–188. New York, NY: Bloomsbury.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18. https://doi.org/10.1515/ling-2017-0032.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Le Foll, Elen. 2021. Introducing the Multi-Feature Tagger of English (MFTE). Perl. Osnabrück University. https://github.com/elenlefoll/MultiFeatureTaggerEnglish. (5 January, 2022).

Introducing a New Open-Source Corpus-Linguistic Tool: The Multi-Feature Tagger of English (MFTE)

Elen Le Foll and Muhammad Shakir

This poster describes and showcases the latest version of the new Multi-Feature Tagger of English (MFTE). Whist many other use-cases may be envisaged, the MFTE was originally designed for the multi-feature/multi-dimensional analysis (MDA; Biber 1988; 1995; Berber Sardinha & Biber 2014) of register variation in standard English (Le Foll 2021). The most popular taggers for English MDAs are currently the Biber Tagger (Biber 1988) and the MAT (Nini 2014; 2019) (Goulart & Wood 2021: 17).

The MFTE offers a more transparent and easily adaptable open-source alternative. Indeed, at the time of writing, the Biber Tagger is not freely accessible to all and the MAT, whilst released under an open-source license since 2020, aims to replicate the Biber Tagger in its 1988 version. Its GUI does not allow researchers to easily examine its inner workings or to adapt it to their needs. The MFTE, by contrast, is available both as a richly annotated python script and as an executable file and is thus also accessible to linguists with no coding skills.

The extensive documentation includes step-by-step instructions to install the Stanford Tagger (Toutanova et al. 2003) used for the underlying POS-tagging and a detailed tabular list of the MFTE's features and their operationalisations. Originally written in Perl (Wall 1994), this new Python 3 (Van Rossum & Drake 2009) version relies on a more widely used language, thus enabling more (corpus) linguists to use the tool and adapt it to their needs. The default tagset comprises 78 lexico-grammatical features ranging from WH-questions to emojis and emoticons. The new python MFTE also features an optional extended tagset which allows for the tagging of an additional 64 features, including many semantic features such as human nouns and verbs of causation (mostly based on dictionary lists from Biber 2006).

The MFTE outputs three comma-separated tables:

1. A table of raw frequency counts

2. A table of frequency counts normalised per 100 words

3. A table of frequency counts with three different normalisations (per 100 finite verb phrases, per 100 nouns and per 100 words) for different types of linguistic features.

We also report on the preliminary results of a formal evaluation of the MFTE's output on a British, American, and South Asian English texts from a wide range of registers.

We hope that the use of a more transparent, open-source tool with detailed evaluation results for each feature (including recall and precision) will contribute to improving the reproducibility and replicability of MDAs.

References

Berber Sardinha, Tony & Douglas Biber (eds.). 2014. *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber* (Studies in Corpus Linguistics (SCL) 60). Amsterdam: John Benjamins.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of Register Variation*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. *University language: a corpus-based study of spoken and written registers* (Studies in Corpus Linguistics v. 23). Amsterdam: John Benjamins.

Evert, Stephanie. 2018. Statistics for Linguists with R – A SIGIL Course: Unit 7: A multivariate approach to linguistic variation. FAU Erlangen-Nürnberg. http://www.stephanieevert.de/SIGIL/sigil_R/. (4 November, 2021).

Goulart, Larissa & Margaret Wood. 2021. Methodological synthesis of research using multidimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science* 6(2). 107–137. https://doi.org/10.1558/jrds.18454.

Le Foll, Elen. 2021. Introducing the Multi-Feature Tagger of English (MFTE). Perl. https://github.com/elenlefoll/MultiFeatureTaggerEnglish. (5 January, 2022).

Nini, Andrea. 2014. Multidimensional Analysis Tagger (MAT). http://sites.google.com/site/multidimensionaltagger. (18 September, 2019).

Nini, Andrea. 2019. The Multi-Dimensional Analysis Tagger. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 67–96. New York: Bloomsbury.

Toutanova, Kristina, Dan Klein, Christopher D. Manning & Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 173–180. Association for Computational Linguistics.

Van Rossum, Guido & Fred L. Drake. 2009. Python 3 Reference Manual. Python Software Foundation. https://docs.python.org/3/reference/. (4 January, 2019).

Wall, Larry. 1994. The PERL Programming Language. http://www.fxjyzy.com:8080/ebook/%E5%B9%BF%E4%BF%A1%E4%B9%A6%E5%BA%93/1211/gjfd/ts005085.pdf. (2 November, 2021).

Composite predicates with *have* and *take* in Late Modern English: a corpus-based investigation

Ljubica Leone

Composite predicates (hereafter CPs) are phraseological verbs consisting of a verb (V) and a deverbal noun (N) which works as the "semantic focus" of the whole V + N combination (Algeo 1995: 204). Examples of CPs include verbs like *have a chance*, *make fun*, and *take time* (Biber et al. 2021).

Many studies have examined the morphological status of CPs (Quirk et al. 1985; Algeo 1995) and described their grammatical functions (Algeo 1995; Biber et al. 1999), or their semantic features (Live 1973). Diachronically, existing works have highlighted that CPs "vastly" increased their nominal constituents during the Middle English (ME) period (Brinton & Traugott 2005: 130; see Matsumoto 1999), and moved towards the fixedness of instances up to the Early Modern English (EModE) period (Akimoto & Brinton 1999; Hiltunen 1999; Kytö 1999; Claridge 2000; Matsumoto 2008). However, despite wide knowledge, there are some areas that are open to investigation concerning the development of CPs during the more recent Late Modern English (LModE) period. Indeed, studies with a focus on LModE have investigated CPs with an emphasis on idiom formation and collocations (Akimoto 1999) or described their functional characteristics (Matsumoto 2005, 2007, 2008) without tying with aspects related to the role performed by grammaticalization, lexicalization, and semantic reanalysis in the renewal of CPs.

The present study aims to fill this gap and to provide a description of processes of change and mechanisms affecting CPs during the years 1750-1850. Specifically, the present research focuses on CPs with *have* and *take* which are considered prototypical examples of the whole class of CPs (Live 1973; Kytö 1999; Matsumoto 2005, 2008).

The present research is a corpus-based investigation undertaken on the Late Modern English-Old Bailey Corpus (LModE-OBC), a corpus that has been compiled by selecting texts from the Proceedings of the Old Bailey (https://www.oldbaileyonline.org/), London's Central Criminal court. The corpus includes trials and witness depositions dating back to the years 1750-1850 and overall amounts to 1,008,234 words. CPs with *have* and *take* have been examined with the software WordSmith Tools 6.0 and specifically with the tool 'Concord' which allows concordance-based analysis and the visualization of CPs and their immediate context.

The analysis reveals that both stability and change characterise CPs with *have* and *take* during the years 1750-1850. New instances established as the result of grammaticalization and lexicalization. Some instances were affected by semantic renewal.

References

Akimoto, Minoji & Laurel J. Brinton. 1999. The origin of the composite predicates in Old English. In Brinton, Laurel J. and Minoji Akimoto (eds.), *Collocational and idiomatic aspects of composite predicates in the history of English*, 21–58. Amsterdam/Philadelphia: John Benjamins.

Akimoto, Minoji. 1999. Collocations and idioms in Late Modern English. In Brinton, Laurel J. and Minoji Akimoto (eds.), *Collocational and idiomatic aspects of composite predicates in the history of English*, 207–238. Amsterdam/Philadelphia: John Benjamins.

Algeo, John. 1995. Have a look at the expanded predicate. In Bas Aarts and Charles F. Meyer (eds.), *The verb in contemporary English: theory and description*, 203–17. Cambridge: Cambridge University Press.

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

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 2021. *Grammar of spoken and written English*. Amsterdam/Philadelphia: John Benjamins.

Brinton, Laurel J. & Elizabeth Closs Traugott. 2005. *Lexicalization and language change*. Cambridge: Cambridge University Press.

Claridge, Claudia. 2000. *Multi-word verbs in Early Modern English: a corpus-based study*. Amsterdam: Rodopi.

Hiltunen, Risto. 1999. Verbal phrases and phrasal verbs in Early Modern English. In Brinton, Laurel J. and Minoji Akimoto (eds.), *Collocational and idiomatic aspects of composite predicates in the history of English*, 133–166. Amsterdam/Philadelphia: John Benjamins.

Kytö, Merja. 1999. Collocational and idiomatic aspects of verbs in Early Modern English. In Brinton, Laurel J. and Minoji Akimoto (eds.), *Collocational and idiomatic aspects of composite predicates in the history of English*, 167–206. Amsterdam/Philadelphia: John Benjamins.

Live, Anna H. 1973. The *take-have* phrasal verb in English. *Linguistics* 95. 31–50.

Matsumoto, Meiko. 1999. Composite predicates in Middle English. In Laurel J. Brinton and Minoji Akimoto (eds.), *Collocational and idiomatic aspects of composite predicates in the history of English*, 59–95. Amsterdam/Philadelphia: John Benjamins.

Matsumoto, Meiko. 2005. The historical development and functional characteristics of composite predicates with *have* and *take* in English. *English Studies* 86 (5). 439–56.

Matsumoto, Meiko. 2007. The verbs *have* and *take* in composite predicates and phrasal verbs. *Studia Neophilologica* 79. 159–170.

Matsumoto, Meiko. 2008. *From simple verbs to periphrastic expressions. The historical development of composite predicates, phrasal verbs, and related constructions in English*. Bern: Peter Lang.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Scott, Mike. 2015. *WordSmith Tools*. Version 6.0. Oxford: Oxford University Press.

The Proceedings of the Old Bailey. https://www.oldbaileyonline.org/

Grammaticalization of Aspect in German and its diachronic parallels in English

Zlata Liwschin

German is commonly not viewed as a typical aspect language, for in German the perfective and imperfective aspectual distinctions are not marked morphologically on the verb, as it is common in the Slavic languages, particularly in Russian (□omrie, 1976; Forsyth, 1970; Leiss, 1992). However, aspect is a grammatical category that is currently being discussed as grammaticalizing in German (Gárgyán, 2014; Krause, 2002). While in Old English and Old High German participial constructions with *beon/wesan* and *sin/wesan* respectively exhibited the primary function of creating internal temporal constituency (Reimann, 1997), upholding the aspectual function and gradually extending their combinability with certain verb classes, in their subsequent stages the initial similarities between the two languages developed apart: In Early Modern English the progressive form gradually became an obligatory member in the English verbal paradigm, whereas in Early New High German the durative function of the construction eventually ceased to exist, this development leading to the disappearance of the form in the German language after the 15th century (Reichmann & Wegera, 1993).

In view of the historical developments, the present work endeavours to define the similarities between the two languages German and English in the way they grammaticalize the category of aspect by acquiring progressive aspect forms. While the English progressive is fully grammaticalized, German progressive constructions are lagging behind, but – as stated by Reimann (1997) (and many others) - German is on its way towards developing obligatory, i.e. fully grammaticalized progressive aspect marking. The study strives to uncover the similarities and differences between the progressives in the two languages in the way they emerged. It will be investigated whether the grammaticalization process in the English language resembles the supposedly presently emerging, presumably similar process in Modern Standard German. For instance, a striking parallel exists between the German *am*- and *beim*-progressives and the Early Modern English locative constructions of the type ‘be in hunting’, built also with the prepositions ‘on’, ‘at’, or ‘upon’ (Núñez-Pertejo, 2004), showing a close formal parallel to the Modern German prepositional progressive forms.

To this end, a comparative corpus study of the progressive in Early/Late Modern English as well as of the progressive forms in Present-Day German is conducted that draws on grammaticalization theory (Lehmann, 2015; Diewald & Smirnova, 2012) as well as on aspectual theory (□omrie, 1976; Leiss, 1992; Bache, 1985). For the English part, the current version of the ARCHER corpus is used, and for the investigation of Present-Day German, DWDS corpus data are analyzed, with a focus on conceptually near-spoken register.

The analysis of corpus data indicates that the internal temporal constituency of a situation is increasingly expressed obligatorily by the *am*-Progressive of the type “Gitarrenmusik ***ist am aussterben***”. Furthermore, my research shows that this obligatory expression can particularly be associated with certain lexical aspect classes. In my ongoing research, I investigate the extent to which the *am*-Progressive behaves syntactically as well as semantically similar to the English Progressive before its complete grammaticalization in the Late Modern English period, such that it may be concluded that both Germanic languages within their relevant stages of diachronic development undergo or underwent a very similar process of grammaticalization of progressive markers, yet at different times in their individual histories.

References

Bache, C. (1985). *Verbal Aspect. A General Theory and its Application to Present-Day English*. Odense: Odense University Press.

Comrie, B. (1976). *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: Cambridge University Press.

Diewald, G., & Smirnova, E. (2012). Paradigmatic integration: the fourth stage in an expanded grammaticalization scenario. In: K. Davidse, T. Breban, L. Brems & T. Mortelmans (Eds.), *Grammaticalization and Language Change. New reflections.* Amsterdam: Benjamins, 111-133.

Forsyth, J. (1970). *A Grammar of Aspect: Usage and Meaning in the Russian Verb.* Cambridge: Cambridge University Press.

Gárgyán, G. (2014). *Der am-Progressiv im heutigen Deutsch: Neue Erkenntnisse mit besonderer Hinsicht auf die Sprachgeschichte, die Aspektualität und den kontrastiven Vergleich mit dem Ungarischen.* Frankfurt am Main: Peter Lang Edition.

Krause, O. (2002). *Progressiv im Deutschen: Eine empirische Untersuchung im Kontrast mit Niederländisch und Englisch*. Tübingen: Niemeyer.

Lehmann, C. (2015). *Thoughts on grammaticalization*. Berlin: Language Science Press.

Leiss, E. (1992). *Die Verbalkategorien des Deutschen: Ein Beitrag zur Theorie der sprachlichen Kategorisierung*. Berlin: Walter de Gruyter.

Núñez-Pertejo, P. (2004). *The progressive in the history of English with special reference to the Early Modern English period: A corpus-based study*. München: LINCOM Europa.

Reichmann, O., & Wegera, K. (1993). *Frühneuhochdeutsche Grammatik* (=Sammlung kurzer Grammatiken germanischer Dialekte. A. Hauptreihe Nr. 12). Tübingen: Niemeyer.

Reimann, A. (1997). *Die Verlaufsform im Deutschen und Englischen. Entwickelt das Deutsche eine „progressive form"?* Bamberg: Dissertation.

Crisis communication across spaces: How COVID-19 was told in World Englishes

Lucía Loureiro-Porto

Socio-cultural and linguistic changes often occur together. For example, Myhill (1995) found that at the time of the American Civil War changes in socio-cultural norms fostered solidarity and the avoidance of unequal power markers, this observed at the linguistic level in a reduction in the frequency of certain modal verbs expressing strong deontic meanings (e.g. *shall, must*). On similar lines, Baker (2010) noted how the shift in society towards less gender-biased attitudes in the second half of the 20th century had an influence on language (i.e. a decreased use of gendered pronouns in the Brown family of corpora). On a related issue, social crises are often accompanied by specific communication techniques, especially in the media, these seen in the power relations between writers and readers, as described by Seoane and Loureiro-Porto (forthcoming). With the aim of exploring the relation between crisis communication and language variation, this paper considers the linguistic effects of the coronavirus pandemic. Existing studies here have shown particular discourse dynamics in terms of the use of attitudinal markers (Dong, Buckingham and Wu 2021), as well as stance nouns (Curry and Pérez-Paredes 2021). My focus is on the use of assertive and nonassertive linguistic markers over time in several Inner Circle and Outer Circle varieties of English (GB, US, NZ, SG and ZA), as represented in the Coronavirus Corpus (Davies 2019-). My working hypothesis is that crisis communication requires an unambiguous style, and hence, under urgent and critical circumstances the frequency of assertive markers increases, while that of non-assertive markers decreases. The variables, then, include assertive markers, as found in Biber's (1988) Dimension 4 ('overt expression of persuasion'), such as suasive verbs, conditional subordination and necessity modals, plus non-assertive markers, from Biber's (1988) Factor 7, which include hedging strategies, downtoners and concessive subordination. From my findings different pictures emerge for the five varieties considered, these in agreement with (i) the different timings and intensities of the COVID-19 waves in the different territories, and also (ii) the differing measures imposed in each country to reduce the spread of the virus. Very interestingly, the results show that the classical division between Inner and Outer-Circle varieties of English does not seem to be at work in the expression of persuasion; what seems to be the main factor in the kind of communication used during this crisis is that of localized strategies adopted to stop the spread of the virus.

References

Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press: Edinburgh.

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

Curry, Niall & Pascual 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.

Davies, Mark. 2019-. *The Coronavirus Corpus*. Available online at https://www.english-corpora.org/corona/.

Dong, Jihua, Louisa Buckingham & Hao Wu. 2021. A discourse dynamics exploration of attitudinal responses towards COVID-19 in academia and media. *International Journal of Corpus Linguistics* 26(4): 532–556.

Hilpert, Martin. 2020. The great temptation. What diachronic corpora do and do not reveal about social change. In Paula Rautionaho, Arja Nurmi & Juhani Klemola (eds.), *Corpora and the Changing Society: Studies in the Evolution of English*. Amsterdam: Benjamins, pp. 3–28.

Myhill, John. 1995. Change and Continuity in the Functions of the American English Modals. *Linguistics* 33 (2): 157–211.

Seoane, Elena & Lucía Loureiro-Porto. Forthcoming. A diachronic corpus-pragmatic approach to the intersection of democratization, colloquialization and popularization: The evolution of newspaper editorials in 1860-1970. Special Issue Corpus-pragmatic studies of democratization in public discourses, ed. by Turo Hiltunen, Turo Vartiainen & Jenni Räikkönen in *Journal of Historical Pragmatics*.

Subtle Differences between the *Help-V* and *Help-to-V* Sequences [Word in progress report]

Noriko Matsumoto

The *help-V* sequence is semantically similar to the *help-to-V* sequence, as in (1). We call the two sequences in (1) 'semantically competing sequences'.

(1) a. He helped organize the party.

b. He helped to organize the party.

Some studies (e.g., McEnery & Xiao 2005) show that the difference between the two semantically competing sequences is based on the one between American and British English. Based on data from the *Collins Wordbanks Online* (CWO), this talk examines whether the two semantically competing sequences allow the distinction between American and British English. It also shows subtle differences between the *help*-V and *help-to*-V sequences.

We use the American and British subcorpora in the CWO. The American and British subcorpora are also divided into the four genres, the book, ephemera, magazine, and broadcast genres. There are four main findings from our corpus data. First, with respect to the verb selection, the *help-V* and *help-to-V* sequences in British English indicate the similar distribution, but the ones in American English do not. Second, with respect to the frequency of use per million words, in American English, the *help-V* sequence shows a much higher frequency than the *help-to-V* sequences. In British English, the *help-V* sequence shows a slightly higher frequency than the *help-to-V* sequence. Third, with respect to genres, each sequence in American and British English shows the relatively similar distribution. In the *help-V* sequence in American and British English, the ephemera genre shows the highest frequency. Third, with respect to inflectional categories of the verb *help*, both in American and British English, the non-past *help-V* sequence shows a much higher frequency than the non-past form *help-to-V* sequence. However, both in American and British English, the *helps/helped/helping-to-V* sequence shows a higher frequency than the *helps/helped/helping-V* sequence. Fourth, both in American and British English, the non-past *help-to-V* sequences following auxiliary verbs show the highest frequency. In American English, the non-past *help-V* sequence sequences following auxiliary verbs shows the highest frequency, while in British English, the non-past help-V sequence following *to*-infinitives shows the highest frequency.

From the above discussion, the difference between American and British English is not the crucial factor differentiating between the *help*-V and *help-to*-V sequences. There is always the possibility that the differences between the two semantically competing sequences are closely related to the verb selection, inflectional categories, grammatical categories followed by the verb help, and genres in a complicated way.

Reference

McEnery, A. and Z. Xiao. 2005. 'HELP or HELP to: What do corpora have to say?' *English Studies*, Vol.86, No.2: 161-187.

Grammatical change in English as a lingua franca: multifactorial analysis of modal verbs with Bayesian modelling

Chunyuan Nie

Multifactorial study on grammatical change in English as a lingua franca (ELF) is still at its preliminary stage while extensive in native (L1) and second language (L2) varieties of English. This study intends to fill this gap by investigating grammatical change in modal verbs of obligation and necessity in ELF with a multifactorial analysis using the Bayesian approach.

The English core modal must and semi-modal have to are undergoing frequency shifts in both L1 and L2 varieties, with must decreasing, and have to is increasing and substituting must in expressing obligation and necessity (Krug 2000; Leech 2003; Collins 2009). The results of multivariate analysis in L1 and L2 varieties indicate that both linguistic and social factors influence this alternation between the two variants (Tagliamonte 2004; Tagliamonte & D'Arcy 2007; Hansen 2018).

This study extracts the two independent variables (must and have to) from two parallel ELF corpora: The Vienna-Oxford International Corpus of English (VOICE 2013) and The Asian Corpus of English (ACE 2020). While VOICE focuses on European speakers, ACE includes Asian speakers. The rationale for including the two corpora is to explore possible regional and typological differences in ELF use. This study exploits information of speakers (age, gender and first language) from corpus metadata and integrates these social factors with linguistic factors in the multifactorial analysis. Instead of the frequentist approach, this study follows the Bayesian approach to inference which allows discussing subtle differences between datasets (Levshina 2022). Bayesian statistics have not yet been used in the study of ELF. This study sets out to answer two research questions: (1) To what extent do ELF speakers follow the grammatical change attested in L1 and L2 English varieties? (2) To what extent do Asian and European speakers show (dis-)similarities in the process of grammatical change?

The results of the Bayesian logistic regression analysis show that, firstly, ELF speakers are sensitive to the grammatical change attested in L1 and L2 varieties. Secondly, while Asian and European speakers show similar patterns in the process of grammatical change, they also show distinctive patterns of their own. This study indicates that the grammatical change in ELF is systematic and detectable. In addition, corpus metadata, which has been underused in earlier ELF research, improves our understanding of the dynamics underlying the grammatical change.

Verb Complementation Patterns in African Englishes: A Corpus-based Study

Folajimi Oyebola

An important aspect of grammatical variation in postcolonial Englishes, especially when they are in the third (nativisation) phase of Schneider's Dynamic Model, is innovative patterns of complementation in individual or semantic groups of verbs (Schneider 2007:88). The more advanced a variety is in the developmental process, the more dissimilar its verb complementation patterns are likely to be from those in native English varieties (see e.g. Olavarría de Ersson and Shaw 2003, Mukherjee and Hoffmann 2006, Mukherjee and Gries 2009). This has been attributed to the possible changes triggered by second-language influence and/or typological factors (Partridge 2019:8) in favour of specific verb constructions. In this study, I investigate the verb complementation patterns in Ghanaian, Nigerian and Kenyan English and whether such patterns differ from those in the historical input variety of British English. The choice of these three varieties is motivated by the fact that the existing literature (e.g. Schmied 2008, Jowitt 2019) has identified innovative patterns of verb complementation as one of the grammatical features of African Englishes. For instance, Jowitt (2019:89) notes that verbs such as *give* and *bring* usually produce the pattern PP + NP in Nigerian English (as in 'You will buy for us water'), in contrast to the choice that native English has between NP1 + NP2 ('You will buy us water') and NP1 + *to*- or *for*- PP ('You will buy water for us'; PP= Prepositional phrase; NP=Noun phrase) (see also Akinlotan and Akinmade 2020). This study focuses on the patterns of ditransitive verbs, *that*-complementiser and *to*-infinitive complementiser, and analyses the components of the four English varieties in the International Corpus of English (ICE), Global Web-based English (GloWbE) and News on the Web (NOW) corpora. The preliminary results show some differences and similarities between the English varieties in their choice of complementation patterns. More importantly, there is an indication that the choice of verb complementation patterns in the English varieties depends on the individual verbs involved and is influenced by various cognitive/typological factors (see Schneider 2012, Callies 2016, Kruger & Van Rooy 2016). The findings of the study are then used to discuss the current stage of development (i.e. endonormative orientation, standardisation) of the three African English varieties.

References

Akinlotan, M. & Akinmade, A. (2020). Dative Alternation in Nigerian English: A Corpus-based Approach. *Glottotheory* 10(1-2): 103–125.

Callies, M. (2016). Towards a Process-oriented Approach to Comparing EFL and ESL varieties: A Corpus-study of Lexical Innovations. *International Journal of Learner Corpus Research* 2(2): 229-250.

Jowitt, D. (2019). *Nigerian English.* Berlin: Mouton de Gruyter.

Kruger, H. & Van Rooy, B. (2016). Constrained Language: A Multidimensional Analysis of Translated English and a Non-native Indigenised Variety of English. *English World-Wide* 37(1): 26-57.

Mukherjee, J. & Gries, S. T. (2009). Collustructional Nativisation in New Englishes: Verb Construction Associations in the International Corpus of English. *English World-Wide* 30(1): 27-51.

Mukherjee, J. & Hoffmann, S. (2006). Describing Verb-complementation Profiles of New Englishes: A Pilot Study of Indian English. *English World-Wide* 27(2): 147-173.

Olavarría de Ersson, E. & Shaw, P. (2003). Verb Complementation Patterns in Indian Standard English. *English World-Wide* 24(1): 137-161.

Partridge, M. (2019). *Verb Complementation Patterns in Black South African English*. PhD Thesis, North-West University.

Schmied, J. J. (2008). East African English (Kenya, Uganda, Tanzania): Morphology and Syntax. In R. Mesthrie (ed.), *Varieties of English: Africa, South and Southeast Asia*, pp. 451-471. Berlin: Mouton de Gruyter.

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

Dative alternation in Black South African English and White South African English

Maristi Partridge

The dative alternation and the motivation behind the choices that speakers make has been investigated in several varieties of English over the last couple of years. This includes varieties such as Old English (De Cuypere, 2015), Late Modern English (Wolk et al., 2013), Indian English (De Cuypere & Verbeke, 2013), and other South Asian English varieties (Bernaisch et al., 2014). This feature has not been investigated in varieties of South African English yet. Consequently, this paper investigates the cross-varietal differences and similarities in terms of the constraints that play a role in the dative alternation found in Black South African English (BSAfE) – a second-language (L2) variety of English – and White South African English (WSAfE) – a first-language (L1) variety of English.

The BSAfE and WSAfE data were obtained from a corpus compiled as part of a larger project investigating the role of editorial intervention in South African varieties of English (Kruger & Van Rooy, 2016, 2017). Concordance lines were drawn from each corpus using a list of verbs known to occur in the dative alternation construction (cf. De Cuypere & Verbeke, 2013). These concordance lines were used to annotate the potential predictor variables of the dative alternation (cf., Bernaisch et al., 2014; Bresnan & Ford, 2010; De Cuypere & Verbeke, 2013). The data obtained from these annotations were used to statistically model the choices that speakers make between the double-object construction and the prepositional dative construction.

The results indicate that whereas the predictor variable VARIETY does not play a significant role in the choices that speakers make, the predictor variables indicative of processing strain does play a significant role in the choice between the double-object dative construction and the prepositional dative construction.

References

Bernaisch, T., Gries, S.T. & Mukherjee, J. 2014. The dative alternation in South Asian English(es): modelling predictors and predicting prototypes. *English World-Wide*, 35(1):7-31. https://doi.org/10.1075/eww.35.1.02ber

Bresnan, J. & Ford, M. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, 86(1):168-213. https://doi.org/10.1353/lan.0.0189.

De Cuypere, L. 2015. A multivariate analysis of the Old English ACC+DAT double object alternation. *Corpus Linguistics and Linguistic Theory*, 11(2):225-254. https://doi.org/10.1515/cllt-2014-0011

De Cuypere, L. & Verbeke, S. 2013. Dative alternation in Indian English: A corpus-based analysis. *World Englishes*, 32(2):169-184. https://doi.org/10.1111/weng.12017

Kruger, H. & Van Rooy, B. 2016. The innovative progressive aspect of Black South African English: the role of language proficiency and normative processes. *International Journal of Learner Corpus Research*, 2(2):205-228. https://doi.org/10.1075/ijlcr.2.2.04van

Kruger, H. & Van Rooy, B. 2017. Editorial practice and the progressive in Black South African English. *World Englishes*, 36(1):20-41. https://doi.org/10.1111/weng.12202

Wolk, C., Bresnan, J., Rosenbach, A. & Szmrecsanyi, B. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*, 30(3):382-419. https://doi.org/10.1075/dia.30.3.04wol

“Only fools don’t change their mind(s)” – plural numerical concord and free variation across selected World Englishes

Karolina Rudnicka

Free variation can manifest itself as the availability of two or more options none of which can be singled out as clearly the most appropriate in a given situation (Cappelle, 2009). For Brown & Miller (2013) free variation is variation in which forms can be used without any change or contrast of meaning. Our recent study (Rudnicka & Klégr *accepted*) focuses on free variation with regard to non-verbal plural concord in phrases such as *lose one’s life* in British and American English, compare (1) and (2).

(1) Many people lost their lives.

(2) Many people lost their life.

In particular, we argue that a certain amount of cases such as (1) and (2) can be seen as instances of free variation, despite the general preference of English towards the agreement in number between (plural) subject and object (Quirk, 1985).

The global spread of English and the existence of many different English varieties make research on these varieties a natural step for scholars investigating change and variation in this language. Also, the fact that West-African Englishes seem to differ with regard to the cut-off point on the count/non-count axis from e.g. British English (Sey, 1973; cf. Mesthrie & Bhatt, 2008) makes the investigation of non-verbal plural number concord in the context of free variation a worthwhile task. To the best of our knowledge, there are no other studies that look at this problem. The present paper aims to reach the following objectives:

i) to provide the general statistics of occurrence of selected constructions (*change one’s mind*, *find one’s way*, *learn one’s lesson*, *lose one’s life*) with regard to the presence or lack of non-verbal number concord between subject and object in three outer-circle[6] varieties, namely Ghanaian English, Nigerian English, and Singapore English and one inner-circle variety – British English;

ii) to compare the acceptability of the distributive plural and the distributive singular variants by the language users;

iii) to answer the question of whether we can talk about a variety-specific free variation.

Objective i) is achieved by means of a corpus-based analysis (conducted with data from GloWbE), whereas objective ii) is accomplished with the use of online-based acceptability ratings.

References

Brown, K. & Miller, J. 2013. *The Cambridge Dictionary of Linguistics*. Cambridge: Cambridge University Press.

Cappelle, B. 2009. Can we factor out free choice? In *Describing and Modeling Variation in Grammar*, Dufter et al. (eds.) 183-201. Berlin and New York: Mouton de Gruyter.

[6] The terminology from Kachru's (1988) model of World Englishes is used here.

Davies, M. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). Available online at https://corpus.byu.edu/glowbe/.

Kachru, B. B. 1988. The sacred cows of English. *English Today* 16: 3-8.

Mesthrie, R., & Bhatt, R. M. 2008. *World Englishes: The study of new linguistic varieties*. Cambridge, UK: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. Edinburgh: Longman.

Rudnicka, K. & Klégr, A. *accepted*. Non-verbal plural number agreement. Between the distributive plural and singular: blocking factors and free variation.

Sey, K. A. 1973. *Ghanaian English*. London: Macmillan.

Frequency structure and speech planning in turns-at-talk

Christoph Rühlemann

Corpus linguists are well versed in mining and evaluating word frequencies, which are key to computing collocations, colligations, collostructions, n-grams, semantic prosodies, or semantic associations. They need not be convinced that "strictly speaking at last, the only thing corpora can provide is information on frequencies" (Gries 2009: 11). What frequencies 'mean' in the context of lexis-lexis and text-lexis co-occurrence phenomena is well researched. What frequencies mean in the context of turn-taking has, by contrast, been neglected—in corpus linguistics and beyond. The turn-taking context is not just any context: Taking turns at talk in face-to-face conversation is the prime ecological niche for language use "where the bulk of language usage occurs" (e.g., Holler & Levinson 2019: 639) and the turn is "very likely the basic form of organization for talk-in-interaction" (Schegloff 2001: 230). This study aims to examine patterns of frequency in the context of turn-taking. It will suggest that if word frequencies are brought to examined in the ecological niche in which they occur—turns-at-talk in conversation—insights can be gained into turn design and turn processing. As introspectve data are normally unavailable to researchers in corpus linguistics, the door to investigating how speech is processed in turns-at-talk has been largely blocked. In this study, which is based not only on the BNC but also a novel corpus, the Freiburg Multimodal Interatcion Corpus (FreMIC), such introspective data are available in large amounts in the form of data on pupil size changes during talk for both speakers and recipients. As pupil changes may indicate cognitive intensity, this data then permits us to enquire into the potential relation between, on the one hand, how frequency patterns across and within turns and, on the other, the cognitive costs that such patterning may incur. The analysis will present two novel findings: it will show that speakers order the words they use in the turn based on the words' frequency in such a way that there is an anticlimactic decrease in frequency across turns of varying sizes and it will show that that anticlimax in frequency is inversely correlated with a climactic increase in cognitive intensity. It will conclude by considering implications of the two findings for theories of speech processing and information packaging. Crucially, the study proposes the notion that the intensifying effort is owed to the achievement of an information climax, which entails a progression from high-frequent, low-cost words through intermediate levels to low-frequent, high-cost words. That achievement is beneficial both for the speaker and the recipient as the required cognitive effort builds up gradually.

Semantic transfer, semantic change and semantic prosody

Mathias Russnes

This paper investigates semantic prosody in a diachronic perspective. Although prosodies have been shown to change over time (e.g. Morley & Partington, 2009; Smith & Nordquist, 2012), there is no consensus regarding the source of such changes; does a semantic change have to occur first, or is the semantic change, and thereby the prosody, triggered by semantic transfer induced by persistent prosodies over time, as some scholars have claimed (see Stewart, 2010)?

The difficulty of separating the process of semantic transfer from other kinds of linguistic change has previously been emphasised by Helsper & Hoffmann (2016). This paper aims to explore this further through a corpus study of the development of the verb lemma FABRICATE, from the late 15th century to the late 20th century, drawing on material from Early English Books Online (EEBO, 1475-1700), the Corpus of Late Modern English Texts (CLMET, 1720-1920), and the British National Corpus 1994 (BNC1994, late 20th century).

A preliminary analysis using MI score and manual scrutiny, following Sinclair's extended-unit-of-meaning model (Sinclair, 1996), suggests that prosodies change over time, and that this change may (in part) be caused by a process of semantic change. The change in this instance seems to be triggered by the emergence of a second meaning ('make up'; to frame or 'invent' (OED)) through a process of metaphorical extension (Allan 2008), rather than through semantic transfer, with a markedly different prosody from the original meaning, synonymous with 'produce'. In EEBO, the original meaning accounts for 85% of the instances, predominantly occurring in positive contexts, e.g. (1). In the CLMET, however, the second, metaphorical meaning makes up around 50% of the occurrences, all of which convey negative evaluative meaning, e.g. (2). The material from BNC1994 offers a similar picture to that of the CLMET in that the metaphorical meaning occurs almost exclusively in negative contexts. However, the original meaning's prosody has shifted from positive to neutral, e.g. (3).

(1) That good emperor took a great delight to *fabricate* and build great works (EEBO)

(2) they believed her to be a creature *fabricated* by my over-heated brain (CLEMET3.1)

(3) Fujitsu has been *fabricating* under license from Vitesse Semiconductor Corp (BNC1994)

This paper aims to categorise these prosodic changes in more detail, addressing the following research questions:

(A) Can semantic prosody be caused by a diachronic process of semantic transfer?

(B) Can this process be separated from semantic change?

References

Allan, K. 2008. *Metaphor and Metonymy: A Diachronic Approach*. Chichester: Wiley-Blackwell.

Helsper, D. & Hoffmann, S. 2016. A diachronic study of semantic prosody. Paper presented at ICAME 37, Hong Kong. Morley, J. & Partington, A. 2009. A few *Frequently Asked Questions* about semantic – or *evaluative* – prosody. *International Journal of Corpus Linguistics* 14 (2): 139-158. Amsterdam: John Benjamins Publishing Company.

Oxford English Dictionary Online. 2022. Oxford University Press. fabricate, v. : Oxford English Dictionary (uio.no) [accessed 18 December 2022]

Sinclair, J. 1996. “The search for units of meaning”. Reprinted in J. Sinclair & R. Carter (eds.), *Trust the Text* (2004), pp. 24-48. London: Routledge.

Smith, K. A. & Nordquist, D. 2012. A critical and historical investigation into semantic prosody. *Journal of Historical Pragmatics* 13 (2): 291-312. John Benjamins Publishing Company.

Stewart, D. 2010. *Semantic Prosody. A Critical Evaluation.* New York: Routledge.

Sentiments in British COVID-19 Twitter Discourse

Julia Schilling and Robert Fuchs

The COVID-19 pandemic has upended lives around the globe and led to intense public debate. Linguists have quickly begun to document and analyze COVID-19 discourse (Baines et al. 2021; Saraff et al. 2021), but there is as yet no large-scale analysis of the sentiments and discourse patterns that characterize British COVID-19 discourse. We address this research gap through a systematic comparative analysis of public discourse during the COVID-19 pandemic. Using a Big Data approach, we (1) identify several keywords associated with the pandemic and (2) track the sentiment of these keywords over time and across regions.

As social media posts can both provide insight into and influence public perceptions of the COVID-19 pandemic, our analysis focuses on Twitter discourse in the United Kingdom. The starting point of the analysis is a data-driven approach to identify COVID-19-related n-grams keywords for each month of the pandemic, comparing pandemic discourse to pre-pandemic discourse while filtering out seasonal effects (e.g., discussion of snow in January). Our data include material from January 2019 to July 2022, with more than 50 million geotagged tweets from the United Kingdom. Rather than collecting tweets based on a pre-existing list of keywords, we use a data-driven approach to identify COVID-19 related terms for each month of the pandemic based on log likelihood and log ratio. We then assign these keywords to semantic fields such as COVID-19 NAMES (e.g. *Covid-19*, *SARS-CoV-2*), PUBLIC HEALTH INSTRUCTIONS (e.g. *self-isolation*, *quarantine, PPE*), VACCINATION and PEOPLE/INSTITUTIONS (e.g. *NHS, Boris Johnson, Matt Hancock*).

In order to analyse public opinion regarding the pandemic and its management, we conduct a sentiment analysis of these semantic fields over time. Rather than opting for a dictionary-based approach, which is fast but potentially unreliable, we opted for a supervised machine learning approach, specifically neural networks for deep learning in Python. Training (70%) and test data (30%) consisted of a sample of 3000 tweets, annotated independently by three coders for positive, negative and neutral sentiment. Disagreements were resolved by majority vote.

We then built a long short-term memory (LSTM) model using tensorflow, opting for this approach due to its ability to recognize patterns in natural language. Based on the annotated samples, our model achieved an accuracy of 89% for the test data. Preliminary results indicate that the sentiment used in tweets mentioning COVID-19 changed over time, being most negative during the second lockdown.

References

Baines, Annalise; Ittefaq, Muhammad & Mauryne Abwao (2021) “#Scamdemic, #Plandemic, or #Scaredemic: What Parler Social Media Platform Tells Us About COVID-19 Vaccine“. *Vaccines* 9 (421), pp. 1-16.

Saraff, Sweta; Singh, Tushar & Ramakrishna Biswal (2021) “Coronavirus Disease 2019: Exploring Media Portrayals of Public Sentiment on Funerals Using Linguistic Dimensions”. *Frontiers in Psychology* 12:626638.

The language of spirituality and religion in the Twitterspace

Gerold Schneider

"English going places" includes the virtual space, a "wild west" in which English is the dominating language (Poblete et al. 2011), also used by very many non-native speakers, and where innovations, creative language, L1 transfer, but also aggression and abuse are rampant, and deep cultural differences clash and mix globally. We have chosen a domain that is particularly globalized, rich in metaphors and imagery, and close to identity: the topic of religion and it associations.

We study collocations of *religion* and closely related terms. Collocations detect 1) fixed phrases (e.g. adjective noun-pairs, 2) metaphors and similes (e.g. copula constructions, comparisons with 'like' and 'as X as Y'), and 3) associations when large observation windows and word embedding methods are used (Sahlgren 2006, Baroni and Lenci 2010).

As RQ 1, we explore the associations between the related terms of *spirituality, faith* and *religion*. What are their associations an collocations, are they used as synonyms or do we find large semantic differences?

Results suggest that while there are cultural differences, the semantic differences between spirituality and religion are stronger. Religion is closer to negative associations, while spirituality is discussed in a more positive light (Neubert 2016; Kim, Lee and King 2020). Spirituality is portrayed as innocent and the path to happiness, while religion is steeped in scandals and hatred. A summary of the associations of *spirituality*, *faith* and *religion*, their "crossing spaces", are given in Figure 1, using an automatic method to draw conceptual maps (Eve 2020, Schneider 2022).

Figure 1. Associations of *spirituality*, *faith* and *religion* in an automatically created conceptual map

As RQ 2, we address the question whether automatic non-compositionality measures can be used in order to improve the detection of idioms and religious imagery. Non-compositionality is marked by long distances in the semantic space between the participants (Senaldi et al. 2016) – does this also hold for our collocational patterns, and does it facilitate their detection? Preliminary results suggest that non-compositionality is mirrored in the semantic space.

As corpus, we use a collection over 100,000 tweets *spirituality, faith* and *religion* that have been sampled across a year and come from a varied cultural background world-wide.

References

Baroni, Marco and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36, 4, 673-721.

Eve, Martin Paul. 2022. *The Digital Humanities and Literary Studies*. Oxford: Oxford University Press.

Kim, Seong-Hyeon, Narae Lee, and Pamela Ebstyne King. 2020. “Dimensions of Religion and Spirituality: A Longitudinal Topic Modeling Approach.” *Journal for the Scientific Study of Religion* 59 (1): 62–83.

Neubert, Frank. 2016. *Die diskursive Konstitution von Religion*. Springer.

Poblete, Barbara, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. "Do all birds tweet the same?: characterizing twitter around the world". CIKM '11: Proceedings of the 20th ACM international conference on Information and knowledge management, 1025–1030.

Sahlgren, Magnus. 2006. The Word-Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Doctoral Thesis, University of Stockholm.

Schneider, Gerold. 2022. “Medical topics and style from 1500 to 2018”. In Turo Hiltunen and Irma Taavitsainen (eds.) *Corpus pragmatic studies on the history of medical discourse*. Amsterdam: Benjamins. 49-78.

Senaldi, Marco S. G., Gianluca E. Lebani, Alessandro Lenci. 2016. “Lexical Variability and Compositionality: Investigating Idiomaticity with Distributional Semantic Models“, *Proceedings of the 12th Workshop on Multiword Expressions* (MWE 2016): 21-31.

A corpus-based analysis of vowel production of L1-Chinese learners of English

Martin Schweinberger & Rui Yin

While pronunciation poses a challenge for language learners (Gilakjani and Ahmadi, 2011), it is also the most immediate and direct display of linguistic proficiency. Listeners automatically and subconsciously categorize and infer judgments about speakers based on pronunciation (Flege, 1995). In addition, pronunciation is crucial for intelligibility and is affecting real-life opportunities (jobs, partner choice, etc.). To ascertain difficulties faced by L1-Chinese learners of English, this study ombines acoustic phonetics with computational and applied corpus linguistics to analyse and compare the production of the monophthongal vowels of 148 L1-Chinese learners (CHN) and 107 L1-speakers of English (ENS) based on The *International Corpus Network of Asian Learners of English* (ICNALE).

The study focuses on systematic differences in vowel duration between CHN and ENS. Specifically, the study tests if CHN do indeed not adapt their tongue position when producing long and short vowel pairs (/I, i:/ and /U, u:/) as suggested in previous research (see Su-Hyun and Liu 2013 as well as Deterding 2009).

The study uses Bhattacharya coefficients to check if CHN do indeed not lower their tongue when producing short variants (I and U) of short-long vowel pairs. In addition, the study uses mixed-effects linear regression modelling to determine if CHN then exaggerate durational contrasts to compensate for the lack of spectral differentiation (due to similar tongue positions). The analysis finds that CHN extend the duration of all vowels and exaggerate the difference between long and short vowels to compensate for the lack of qualitative differences between short and long vowel pairs. This study represents the first corpus-based acoustic analysis of CHN vowels in spontaneous speech.

References

Deterding, David. 2009. The Formants of Monophthong Vowels in Standard Southern British English Pronunciation. *Journal of the International Phonetic Association* 27 (1-2), 47-55.

Flege, J. E. 1995. Second-language speech learning: theory, findings, and problems. In W. Strange (ed.), *Speech perception and linguistic experience: Issues in cross-linguistic research*, 233-277. York Press.

Gilakjani, A. P., and Ahmadi, M. R., 2011. Why Is Pronunciation So Difficult to Learn?, *English Language Teaching*, 4(3), 74-83.

Su-Hyun, Jina and Chang Liu. 2013. The vowel inherent spectral change of English vowels spoken by native and non-native speakers. *The Journal of the Acoustical Society of America* 133. https://doi.org/10.1121/1.4798620

Patterns of functional variation in South Asian Online Englishes

Muhammad Shakir

The multidimensional analysis framework (e.g. Biber 1988) has been used to describe variation in online registers from the native English-speaking regions (Biber and Egbert 2016) as well as in the subfield of World Englishes (Xiao 2009; Bohmann 2020). The present study builds on this research and aims to describe functional variation in online English communication of four South Asian countries (i.e. Bangladesh, India, Pakistan, and Sri Lanka) in comparison with the two main native varieties of English (i.e. British English and American English). The data includes categories like newspaper comments, tweets, web forums, and general websites (20% texts published on blogs + 80% from other websites that may include blogs too). Additionally, for the four S. Asian countries English text messages are also included. Two MD analyses have been conducted:

In MD1, there are 1400 texts tagged using a modified version of the MFTE tagger (Le Foll 2022) and analysed using principal component analysis in the R programming language (R Core Team 2022). The results consist of five dimensions of variation:

1. Oral versus literate discourse

2. Oral elaboration versus informational concerns

3. Technical explanation versus past orientation

4. Interaction versus technology focus

5. Reported versus other communication

The four South Asian countries are more literate (D1), informational (D2), tech explanatory (D3), interaction focused (D4), and incline towards reported communication (D5). The analysis of South Asian text messages on MD1 shows that Pakistani text messages are significantly more literate as compared to the other three SA countries, mainly due to the inclusion of official communication and other forwarded texts.

In MD2, individual texts (n=1129) from the general websites section have been rated for their situational characteristics following Biber and Egbert (2018, pp. 196-216). A separate principal component analysis has been run on these texts by constructing a combined data table of the situational features (ratings) and the linguistic features (MFTE tags), which results in four dimensions of variation:

1. Oral personal versus literate informational

2. Explanatory/procedural versus narrative discourse

3. Literate elaboration versus other concerns

4. Literate persuasion

The first 2 dimensions of the MD2 are more relevant and explanatory. The results reveal that the SA countries are literate informational on dimension 1 and narrative on dimension 2 as compared to the two native English-speaking countries that show opposite trends.

The overall results indicate that the SA countries have more homogeneity when compared to the UK and USA, though there are also individual differences among them.

References

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

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

Biber, D., & Egbert, J. (2018). *Register variation online* (1st ed.). Cambridge University Press.

Bohmann, A. (2020). *Variation in English worldwide: Registers and global varieties* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108751339

Le Foll, E. (2022). *Textbook English a corpus-based analysis of the language of EFL textbooks used in secondary schools in France, Germany and Spain* [PhD Thesis]. University of Osnabrück.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Xiao, R. (2009). Multidimensional analysis and the study of world Englishes. *World Englishes*, *28*(4), 421–450.

Gathering Internet Memes for a Corpus of Peer Health Discourse [word in progress report]

Laurel Stvan

About the Corpus

We discuss a social media component of CADOH, the Corpus of American Discourses on Health. Much corpus linguistic work on health focuses on the language of medical providers or medical researchers (e.g., Atkinson & Valle 2012). This corpus focuses instead on examining how health information is conveyed among non-specialists. The current version contains written texts (newspaper, magazines), spoken dialog transcripts (fictional excerpts, TV and radio broadcasts, and multi-speaker student focus group discussions) and online data (blogs and comments, and open forum discussions). These are now supplemented by a collection of food and health-focused internet memes.

Memes as A Corpus Component

The genre of the internet meme fits the corpus goals well due to memes reflecting the following characteristics: an informal, vernacular register; use in peer-to-peer communication; occurring in a wide set of social media outlets; fast spreading transmission; and uses that are both general in topic and tightly reflecting the time and setting of their creation.

Meme Collection

A first round of the health meme dataset was collected in the fall of 2018. Covid-related memes were added in 2020-2022. Googling keywords plus *meme* in image searches we collected 200 memes on 13 topics: foods making one fat, low-fat food, how to lose weight, fat, sugar, cholesterol, diet Coke, colonoscopies, catching a cold, flu shots, covid masking, covid vaccines, and social distancing. Metadata collected includes the URL, author, posting date, exact words included, summary of the topic, summary of the image, and a jpeg of each meme.

Collection Issues and Conclusions

The following issues emerged in including meme data in a corpus: a) Thematic collection is more manageable than characterizing all memes, yet reflects many general meme traits (cf. Dynel 2021, Wang & Wen 2015). b) We found advantages in starting with Google images, since it tracks memes sourced via multiple outlets (FB, Twitter, Instagram). c) Concordancing issues include the (often purposely) aberrant spelling, and weighing html scraping vs. OCR to best capture drawn or superimposed text. d) Multimodality effects: because images and phrases repeat, they are often mix-and-matchable. So, impact can come from picture or text (cf. Highfield & Leaver 2016). e) Meme images can convey humor but may also underscore how they are taken as wisdom or common sense. f) Yet, while presented as authoritative, memes can feature both pros and cons of divisive topics (vaccinations, diet foods, catching a cold).

References

Atkinson, Dwight & Ellen Valle. 2012. Corpus Analysis of Scientific and Medical Writing Across Time. *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd.

Dynel, Marta. 2021. COVID-19 memes going viral: On the multiple multimodal voices behind face masks. *Discourse & Society*. SAGE Publications Sage UK: London, England 32(2). 175–195.

Highfield, Tim & Tama Leaver. 2016. Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji. *Communication Research and Practice* 2(1). 47–62.

Wang, William Yang & Miaomiao Wen. 2015. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 355–365.

Singlish in cyberspace: Effects of the Speak Good English Movement (SGEM) campaign

Teh Cher Huey and Laurence Anthony

Singlish is a creole language that combines English with other languages and dialects in Singapore. Due to its potential implications on successful communication in international settings, Singlish has often been deemed controversial, leading to various initiatives to restrict its usage. One such example is the Speak Good English Movement (SGEM), a nation-wide campaign initiated in 2000 by the Singaporean government to promote the use of Standard English among Singaporeans. While some groups have reacted to the campaign positively, critics have long argued that it advocates an elitist form of language that has led to the devaluation of Singlish. Despite these criticisms, the SGEM was recently relaunch in 2019, after close to a decade of inactivity. This paper aims to investigate the campaign's effect on the Singaporean creole language as it is used in cyberspace.

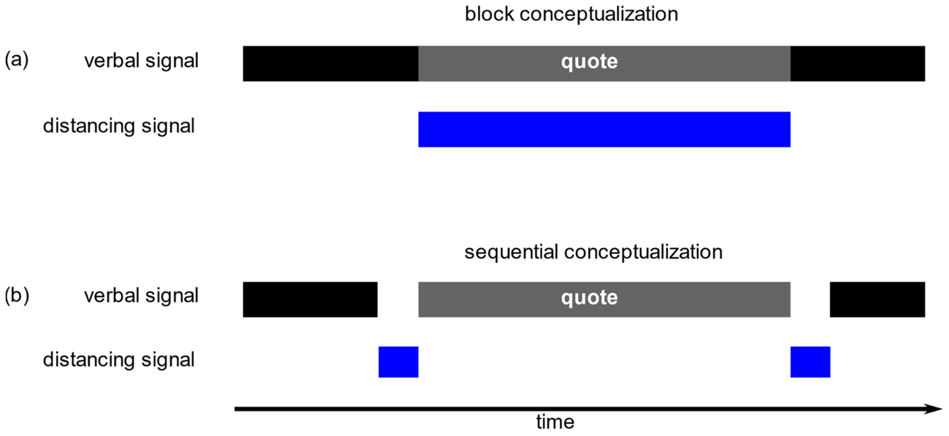
The study focuses on the usage of Singlish sentence-final discourse particles, a linguistic feature often described as the most salient feature of Singlish. Reddit comments from the r/Singapore subreddit were collected from 2018 to 2022 using the Python Reddit API Wrapper. Next, changes in the usage of such discourse particles were investigated using statistical and qualitative corpus methods. Results reveal that the use of Singlish in online discourse varies considerably depending on the individual, but it has remained relatively stable over the period of study despite the relaunch of the SGEM campaign. The study therefore argues that Singlish usage in online spaces reflects a divergent relationship between the Singaporean government's language policies and the language practices of Singaporeans. We anticipate that these findings can contribute to the current literature on Singlish usage and provide valuable insights for language policy makers and educators. The study also demonstrates the potential of corpus linguistic analysis in understanding the use of English in transnational contexts, specifically in the online space.

Forms and Function of Air Quotes in American TV News

Peter Uhrig

The use of so-called air quotes in spoken discourse is a very interesting feature in terms of language evolution, where traditionally the written modality has been regarded as secondary and derived. It is relatively uncommon to see a typographic element find its way into spoken discourse, although we do for instance find sentence-end markers (such as *period* or *full stop*) pronounced to say that something is definitive and constitutes the end of discussion. However, air quotes are special because they occur in the visual channel, although, as this paper will show, they tend to co-occur with certain prosodic features and verbal items such as *quote* or *quote-unquote*.

In this paper, the full range of formal and functional variation found in a large corpus of American TV News (Uhrig 2018) will be discussed against the backdrop of existing literature on the topic (e.g. Stivers & Sidnell 2005, Lampert 2015, Cirillo 2019). It will be shown that there is more variation than often acknowledged, not only in the gestural features (see McNeill 1992 for a classification scheme) but also in the co-occurrence with prosodic cues and verbal elements, and the timings found. With regard to the interplay of these various methods of signaling distance, it will be argued that there are two different modes of conceptualization used by speakers to mark elements enclosed in some form of quotative element, viz. sequential conceptualization and block conceptualization, prototypes of which are illustrated Figure 1:



(1) And so I went to the "place where you get the beer from", and there were four ladies who were serving up, and I stood up[7] [my quotation marks]

Example (1) illustrates an extreme case of block conceptualization, with air quotes and prosodic cues covering the entire passage, whereas the older passage found behind the second QR code (which is also a clickable link) shows a serial conceptualization.

The paper will analyze a range of less prototypical example, some of which show both types of conceptualization in different communicative channels, and suggest possible ways of modelling these occurrences in a Construction Grammar framework.

[7] 2019-05-22_2200_US_KNBC_The_Ellen_DeGeneres_Show.txt; note that the transcriptionist did not put any quotation marks in.

References:

Cirillo, Letizia. 2019. The pragmatics of air quotes in English academic presentations. *Journal of Pragmatics* 142.

Lampert, Martina. 2015. ‘Speaking’ Quotation Marks: Toward a Multimodal Analysis of Quoting Verbatim in English. Frankfurt a.M.: Peter Lang.

McNeill, David. 1992. Hand and Mind: What Gestures Reveal About Thought. Chicago, IL: University of Chicago Press.

Stivers, Tanya & Jack Sidnell. 2005. Introduction: Multimodal interaction. *Semiotica* 156-1/4.

Uhrig, Peter. 2018. NewsScape and the Distributed Little Red Hen Lab – A digital infrastructure for the large-scale analysis of TV broadcasts. In: Anne-Julia Zwierlein, Jochen Petzold, Katharina Böhm and Martin Decker (eds.), *Anglistentag 2017 in Regensburg: Proceedings. Proceedings of the Conference of the German Association of University Teachers of English*. Trier: Wissenschaftlicher Verlag Trier.

A variational metapragmatic study of promising and complaining in West African Englishes: A corpus-based study

Foluke Unuabonah and Ibukun Filani

Previous accounts of meta-illocutionary lexicon have investigated these pragmatic items in different varieties of English such as Irish English, American English and British (Schneider, 2021, 2022; Schoppa, 2021), without exploring them in African varieties of English. In order to close this gap, this paper explores the meta-illocutionary lexicon, COMPLAIN and PROMISE, and their derivational and inflectional variants in two varieties of West African Englishes: Ghanaian English and Nigerian English, from a variational corpus metapragmatic model. In particular, this paper focuses on their frequencies, pragmatic functions as well as their occurrence in different text types. The data, which were quantitatively and qualitatively analysed, were extracted from the Ghanaian and Nigerian components of the International Corpus of English, using AntConc (Anthony, 2015). The results reveal that while there is a significant difference in the frequency of the meta-illocutionary lexicon of PROMISE between the two varieties, there is no significant difference in the frequency of the meta-illocutionary lexicon of COMPLAIN between the two West African varieties. Moreover, there are significant differences in the use of these meta-illocutionary lexicon across different text types between the two West African varieties. In addition, the results show that while PROMISE is used to perform commenting, reporting, performative and clarifying functions, COMPLAIN is used to perform commenting, reporting, performative functions. The results are explained based on linguistic and sociocultural factors. Thus, the paper contributes to the research on meta-illocutionary lexicon, variational metapragmatics, corpus pragmatics, and world Englishes in general.

References

Anthony, L. (2015). AntConc (Version 3.4.4) [Computer software]. Retrieved from https://www.laurenceanthony.net/software.

Schoppa, Dominik J. (2022). Conceptualizing illocutions in context: A variationist perspective on the meta-illocutionary lexicon. *Corpus Pragmatics*, 6, 63–88.

Schneider, Klaus P. (2021). Notes on variational metapragmatics. *Journal of Pragmatics*, 179, 12-18.

Schneider, Klaus P. (2022). Referring to speech acts in communication: Exploring meta-illocutionary expressions in ICE-Ireland. *Corpus Pragmatics*, 6, 155-174.

Speaking of variation and, uhm, complexity in a corpus of spoken dialogues

Thomas Van Hoey, Matt H. Gardner and Benedikt Szmrecsanyi

In this presentation, we investigate the relative complexity (i.e., difficulty) incurred by having to choose between competing grammatical variants. While variational linguists provide overwhelming evidence for the existence, ubiquity, and systematicity of variable patterns — or "alternate ways of saying 'the same' thing" (Labov, 1972: 188), as in *Tom picked up the book* versus *Tom picked the book up* — there is a fairly widespread assumption in linguistics that grammatical variation should create undue complexity for language users. This is because grammatical variation (as opposed to e.g., lexical variation) is typically conditioned probabilistically by any number of contextual constraints. Even before language users can make a choice as a function of the naturalness of a grammatical variant in a specific linguistic context, they need to check that linguistic context for the various constraints that regulate the variation at hand. It follows that this extra cognitive work must increase cognitive load. Or does it?

Against this backdrop, we report on the extension of a study (Gardner et al. 2021) that explores the link between production difficulty and grammatical variability using a corpus-based research design. The idea is that if isomorphism à la Haiman (1985) is a design feature of human languages, then variation — to the extent that it exists — should be suboptimal as it complicates language economy and efficiency for both speaker and addressee. Suboptimality, in turn, should be measurable by quantifying the extent to which variation contexts attract production difficulties.

Contrary to traditional expectations of the field, a multivariate analysis based on the Switchboard Corpus of American English (542 speakers, 240 hours of recording) shows that, on a turn-by-turn basis, the presence of variable contexts does not positively correlate with two metrics of production difficulty, namely filled pauses (*um* and *uh*) and unfilled pauses (speech planning time). When 20 morphosyntactic variables are considered collectively (N = 57,660 choice contexts), there is no significant effect. In other words, choice contexts do not correlate with measurable production difficulties. These results challenge the view that grammatical variability is somehow suboptimal for speakers.

References

Gardner, Matt Hunt, Eva Uffing, Nicholas Van Vaeck & Benedikt Szmrecsanyi. 2021. Variation isn't that hard: Morphosyntactic choice does not predict production difficulty. (Ed.) Stefan Th. Gries. *PLOS ONE* 16(6). e0252602. ttps://doi.org/10.1371/journal.pone.0252602.

Haiman, John. 1980. The Iconicity of Grammar: Isomorphism and Motivation. *Language* 56(3). 515. https://doi.org/10.2307/414448.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.

The curious case of lexicogrammatical change: approaches to language contact

Ronel Wasserman

What actually happens on- and offstage when languages and their speakers come into contact? For features of South African English varieties, language contact has been considered the direct catalyst for lexicogrammatical change (cf. Wasserman 2020), but opposing views also exist in favour of considerations linked to endogenous linguistic development (cf. Lass & Wright 1986; Mesthrie 2002). Apart from arguments related to specific features, and based both on corpus or other kinds of evidence, the more general nature of the role of language contact in shaping languages in contact continues to be examined from various linguistic positions, often without overt exchange. This paper offers a review of various micro- and macrolinguistic approaches to understanding the relationship between language change and language contact, working towards constructing an integrated theoretical framework for (possible contact-induced) lexicogrammatical change in English varieties, and potentially language, in South Africa and beyond.

Contact is considered from, for example, the broader (and occasionally cross-fertilising) perspectives of contact linguistics, evolutionary linguistics, historical sociolinguistics, cognitive contact linguistics, the anthropology of language contact, and the psycholinguistics of language change (e.g. Croft 2000; Nevalainen 2009; Zenner et al. 2019; Weinreich 1953; Myers-Scotton 2002). Perspectives on contact-induced change related to grammaticalisation theory and the relevant branches of construction grammar, such as diachronic construction grammar, constructional contact linguistics, and diasystematic construction grammar, are furthermore compared (e.g. Heine & Kuteva 2005; Traugott & Trousdale 2013; Hilpert 2018; Höder 2018; Boas & Höder 2021). Ultimately, the review illustrates the interaction between the Labovian trifecta of internal, social, and cognitive and cultural factors (e.g. Labov 2010), when considering the role of contact in language change.

References

Boas, H. C. & Höder, S. (2021). Widening the scope: Recent trends in constructional contact linguistics. In H.C. Boas & S. Höder (Eds.), *Constructions in contact 2: Language change, multilingual practices, and additional language acquisition* (pp. 2- 13). John Benjamins.

Croft, William. (2000). *Explaining Language Change: An Evolutionary Approach*. Person Education

Heine, Bernd and Tania Kuteva 2005. *Language Contact and Grammatical Change*. Cambridge: Cambridge University Press.

Hilpert, M. (2018). Three open questions in diachronic construction grammar. In E. Coussé, P. Andersson, & J. Olofsson (Eds.), *Grammaticalization meets construction grammar* (pp. 21–39). John Benjamins.

Höder, S. (2018). Grammar is community-specific: Background and basic concepts of Diasystematic Construction Grammar. In H. C. Boas & S. Höder (Eds.), *Constructions in contact: Constructional perspectives on contact phenomena in Germanic languages* (pp. 37-70). John Benjamins.

Labov, William. 2010. *Principles of Linguistic Change, Cognitive and Cultural Factors. Vol. 3 of Principles of Linguistic Change*. Wiley-Blackwell.

Lass, R. and Wright, S. (1986). Endogeny vs. contact: ‘Afrikaans influence’ on South African English. *English World-Wide*, 7(2), 201–223.

Mesthrie, R. (2002). Endogeny versus contact revisited: aspectual busy in South African English. *Language Sciences*, 24, 345-358.

Myers-Scotton, C. (2002). *Bilingual encounters and grammatical outcomes*. Oxford University Press.

Nevalainen, Terttu. (2009). Historical Sociolinguistics and Language Change. In van A. van Kemenade and L. Bettelou (Eds.), *Handbook of the History of English* (pp. 558-588.). Wiley-Blackwell.

Traugott, E. C. & Trousdale, G. (2013). *Constructionalization and Constructional Changes*. Oxford University Press.

Wasserman, R. (2020). The Historical Development of South African English: Semantic Features. In R. Hickey (Ed.), *English in Multilingual South Africa: The Linguistics of Contact and Change.* (pp. 52–73.) Cambridge University Press.

Weinreich, Uriel. (1953). *Languages in contact: Findings and problems*. Mouton.

Zenner, E., Backus, A. & Winter-Froemel, E. (2019). *Cognitive Contact Linguistics: Placing Usage, Meaning and Mind at the Core of Contact-Induced Variation and Change*. De Gruyter Mouton.

My Astrea, my dearest Miss Goldsworthy, ever dear Lady Wake – Direct forms of address in Mary Hamilton's private correspondence

Nuria Yáñez-Bouza

Eighteenth-century England witnessed significant social and cultural changes through which politeness became an "ideal that was aspired to in all aspects of daily life", and this included language use (Jucker 2020: 117-134). It is also the period when letter writing became a widespread social practice, with a proliferation of letter-writing manuals laying down rules for addressing persons of all ranks with propriety and elegance of style (Bannet 2005). In this context, the use of appropriate forms of address became de rigueur and a key element of socially-governed linguistic practice (Whymann 2009).

Previous studies in the field of historical sociolinguistics and historical sociopragmatics have investigated expressions of address towards the recipient in the Early Modern Engish period (Nevalainen & Raumolin-Brunberg 1995, Nevala 2004) and also in some eighteenth-century letter writers (Tieken-Boon van Ostade 1999, 2011, 2014). The case study presented in this paper aims to build on earlier work by exploring the use of direct address as an index of politeness which reflects the mutual relationship between correspondents, and with a particular focus on intra-speaker variation in the use of personal names and honorific terms as a way to convey different values on the positive-negative politeness continuum (Nevala 2004).

The data are drawn from a set of private correspondence written by Mary Hamilton, a well-educated figure with several interlocking royal, aristocratic, literary, and artistic networks in the late Georgian period (1776-1814, 170 items, c.53,000 words; see The Mary Hamilton Papers). The relevant expressions of direct address have been identified through a process of manual reading and data retrieval based on customised XML mark-up in the opening, body, and subscription part of the letters. The socio-pragmatic analysis considers sociolinguistic factors (gender) as well as notions traditionally connected with pragmatic language use (distance, relative power). The findings reveal that Hamilton makes use of a rich variety of forms on the politeness continuum, with a slightly more frequent use of honorifics and status terms, either as head-nouns (my Ladyship, my dear Madam) or with a first/last name (my dear Lady Charlotte), while personal names often appear in the form of a nickname (my Astrea) or in combination with a title and last name (Mr Walpole). It is observed that intra-speaker variation correlates with social distance (friends, acquaintances) and with relative power (equal relation, inferior-to-superior rank), but not with the recipient's gender.

References

Bannet, Eve Tavor. 2005. *Empire of letters: Letter manuals and transatlantic correspondence, 1680-1820*. Cambridge: Cambridge University Press.

Jucker, Andreas H. 2020. *Politeness in the history of English. From the Middle Ages to the present day.* Cambridge: Cambridge University Press.

Nevala, Minna. 2004. *Address in Early English Correspondence: Its forms and socio-pragmatic functions*. Helsinki: Société Néophilologique.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 1995. Constraints on politeness: The pragmatics of address formulae in early English correspondence. In Andreas H. Jucker (ed.), *Historical pragmatics: Pragmatic developments in the history of English* (Pragmatics and Beyond New Series 35), 541-601. Amsterdam: John Benjamins.

The Mary Hamilton Papers (c.1740-c.1850). Compiled by David Denison, Nuria Yáñez-Bouza, Tino Oudesluijs, Cassandra Ulph, Christine Wallis, Hannah Barker and Sophie Coulombeau, University of Manchester. In progress, 2019-, www.maryhamiltonpapers.alc.manchester.ac.uk

Tieken-Boon van Ostade, Ingrid. 1999. Of formulas and friends: Expressions of politeness in John Gay's letters. In Guy A. J. Tops, Betty Devriendt & Steven Geukens (eds.), *Thinking English grammar. To honour Xavier Dekeyser, Professor Emeritus*, 99-112. Louvain: Peeters.

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

Tieken-Boon van Ostade, Ingrid. 2014. *In search of Jane Austen: The language of the letters*. Oxford: Oxford University Press.

Whyman, Susan. 2009. *The pen and the people: English letter writers, 1660-1800*. Oxford: Oxford University Press.