ICAME 38

24–28 May 2017 | Charles University | Prague

Corpus et Orbis: Interpreting the World through Corpora

Book of Abstracts

Table of Contents

Pre-conference Workshops

Workshop 1

Genderlectal differentiation in the Outer and Expanding Circles? Corpus-linguistic explorations

Workshop convenor:
Tobias Bernaisch (Justus Liebig University Gießen)

Tobias Bernaisch
Justus Liebig University Giessen
Tobias.J.Bernaisch@anglistik.uni-giessen.de

Genderlectal variation with hedges in the Inner and Expanding Circle? Corpus-based explorations of British English and Southeast Asian Englishes

Earlier sociolinguistic research into genderlects repeatedly profiled – though not unequivocally (cf. e.g. Baumann 1976; Schultz et al. 1984) – women as prototypical users of hedges (cf. e.g. Lakoff 1975; Holmes 1984, 1995; Coates 1987, 1989), i.e. linguistic devices such as *I believe* or *probably* removing the definiteness from a statement to potentially save interlocutors' faces. Hedges can be realised in different forms, e.g. via modal auxiliaries (*may*, *might*), modal adverbs (*perhaps*, *possibly*) or discourse markers (*I think*, *I mean*), and fulfil numerous functions such as expressing doubt or showing sensitivity to interlocutors' feelings (cf. Coates 2013). For native users of English, more strictly corpus-linguistic gender-oriented explorations of hedges exist as well (cf. e.g. Murphy 2010) and GENDER is sometimes included as a sociolinguistic factor in studies on second-language users (cf. e.g. Lange 2012), but corpus-linguistic investigations of potential gender-preferential uses with learners are not available. This study thus seeks to corpus-linguistically establish a) whether women or men use more hedges in native-speaker and learner contexts, b) what factors determine which concrete hedge a user chooses in a given communicative context and c) on a more theoretical level to what extent the usage of hedges by learners in given postcolonial settings can be reconciled with the evolutionary stage of the postcolonial habitat (cf. Schneider 2007) concerned.

Eight clausal (e.g. *I think*, *I guess*) and eight non-clausal (e.g. *maybe*, *apparently*) hedges were extracted from texts by native speakers from Great Britain and texts by learners from Hong Kong, the Philippines and Singapore not beyond the level B2 in the *Common European Framework of Reference for Languages* as they are available in ICNALE, the International Corpus Network of Asian Learners of English (Ishikawa

2014). Each of the 1530 examples was annotated for REGION, MODE, SEX of user, AGE of user and OCCUPATION of user and statistically modelled via conditional inference trees (cf. Hothorn & Zeileis 2015; Hothorn et al. 2006) and multinomial log-linear regression (cf. Venables & Ripley 2002).

The data show that men use more hedges in Great Britain and Singapore, while women employ more hedges in Hong Kong and the Philippines. *I think* is the most frequent hedge, but the concrete hedge chosen is determined by REGION, where Singapore is notably different from other territories, MODE, SEX and OCCUPATION, where speakers in the humanities choose other hedges than people with more technologically-oriented jobs. Learners in endonormatively stabilised postcolonial settings as in Singapore can be differentiated more clearly from British English users than learners in less evolutionary advanced habitats such as Hong Kong or the Philippines.

References

Baumann, M. (1976). Two features of women's speech?. In B. L. Dubois & I. Crouch (Eds.), The Sociology of the Languages of American Women. San Antonio, TX: Trinity University, 32–40.

Coates, J. (1987). Epistemic modality and spoken discourse. Transactions of the Philological Society, 85, 110–131.

Coates, J. (1989). Gossip revisited: An analysis of all-female discourse. In J. Coates & D. Cameron (Eds.), Women in their speech communities. London: Longman, 94–122.

Coates, J. (2013). Women, Men and Everyday Talk. London: Palgrave Macmillan.

Holmes, J. (1984). Hedging your bets and sitting on the fence: Some evidence for hedges as support structures. Te Reo, 27, 47–62.

Holmes, J. (1995). Women, Men and Politeness. London: Longman.

Hothorn, T. & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. Journal of Machine Learning Research, 16, 3905–3909.

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15 (3), 651–674.

Ishikawa, S. (2014). Design of the ICNALE-Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), Learner Corpus Studies in Asia and the World, Vol 2. Kobe, Japan: Kobe University, 63–76.

Lakoff, R. (1975). Language and Women's Place. New York: Harper and Row.

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

Murphy, B. (2010). Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk. Amsterdam: John Benjamins.

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

Schultz, K., Briere, J. & Sandler, L. (1984). The use and development of sex-typed language. Psychology of Women Quarterly, 8, 327–336.

Venables, W. N. & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. New York: Springer.

• • •

Robert Fuchs
Hong Kong Baptist University
robert.fuchs.dd@gmail.com

Sociolinguistic variation in intensifier usage in Indian and British English: Gender and language in the Inner and Outer Circle

Previous research on gender and language in varieties of English spoken in Inner Circle countries (where English is spoken as a Native Language) has shown that, all else being the same, female speech tends to differ from male speech in that words belonging to certain semantic categories are used more frequently by one group compared to the other. For example, men tend to use specialised colour terms such as 'mauve' less frequently than basic colour terms such as 'yellow', relative to their frequency in female speech. Another example is the use of intensifiers such as 'very' and 'really', which tend to be used more frequently by women than by men. However, other sociolinguistic factors can be confounding factors, and consistent gender differences are often found only when speakers of the same age and educational background, speaking in contexts of similar formality, are compared. For example, Tagliamonte and Roberts (2005) found *so* to be more frequent with women than with men, but York, Ito and Tagliamonte (2003) found no difference. Conflicting results might be due to the fact that small numbers of both types and tokens were considered in these and other studies. Other factors, such as differences in age between interlocutors, whether groups consist of men or women only or are mixed, and whether speakers are equals or not have not been considered so far in quantitative research. Finally, our knowledge of the sociolinguistic associations of intensifier usage in general, and gender differences in particular, is so far restricted to Inner Circle varieties of English. The present paper will address this research gap by contrasting the sociolinguistics of intensifier usage in Indian and British English.

The data used in this study comes from the spoken parts of the Indian and British components of the International Corpus of English, with data from more than 600 speakers. A total of 29 intensifiers (boosters) were investigated, and the number of tokens used by each speaker, as well as the number of words uttered by each speaker were determined in order to compute relative frequencies for each speaker and, based on this, for groups. This has the advantage that speakers who behave unlike others from the same demographic group do not unduly influence group means through their particularly high or low usage of intensifiers. Linear regression models were then computed in the statistical software package R to determine which sociolinguistic factors significantly influence the relative frequency of intensifiers in spoken Indian and British English.

Adopting this robust speaker-based quantitative approach, gender, group composition (all-female, all-male, mixed), age, formality, the number of interlocutors and their relationship are shown to be significant factors. Results show that, in both IndE and BrE, women use more intensifiers when speaking with other women, but in

mixed groups women do not differ from men in intensifier frequency. Furthermore, women use more intensifiers than men in informal, but fewer in formal situations. More intensifiers are also used if speakers are equals in the social hierarchy, and when group size is small. While age is an important factor in determining intensifier frequency, whether interlocutors differ in age has no significant influence. Rather, age interacts with gender and group composition. While these results indicate that gender differences in intensifier usage are relatively similar in the two varieties, a difference was found in how gender and formality interact. In IndE, the female lead in intensifier frequency was greater than in BrE in private conversations, but in more formal contexts, the male lead in intensifier frequency is greater in BrE.

References

Brown, P. & Levinson, S. C. (1987). Politeness. Some Universals in Language Use. Cambridge: Cambridge University Press.

Bulgin, J., Elford, N., Harding, L., HenleyB., Power S. & Walters, C. (2008). Linguistica Atlantica 29, 101-115.

Ito, R. & Tagliamonte S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. Language in Society 32, 257–279.

Tagliamonte, S. & Roberts, C. 2005. So weird; so cool; so innovative: The use of intensifiers in the television series Friends. American Speech 80(3), 280-300.

Yaguchi, M., Iyeiri, Y. & Baba, Y. (2010). Speech style and gender distinctions in the use of very and real/really: An analysis of the Corpus of Spoken Professional American English. Journal of Pragmatics 42, 585-597.

• • •

Sandra Götz

Justus Liebig University Giessen

Sandra.Goetz@anglistik.uni-giessen.de

Investigating gender-related stereotypes in Jamaican English: A corpus-based approach

Sociolinguists have documented great variation in male vs. female speaking styles, either corresponding to gender-exclusive or gender-preferential features (Holmes 2013). Common linguistic stereotypes in this vein claim that women, in general, talk more than men, women generally use more standard, indirect and polite forms than men, who, in turn, are said to use more vernacular forms accordingly, women using more diminutives, euphemisms, expressive forms, intensifiers, hedges, minimal responses, etc. (e.g. Tannen 1994; Lakoff 2004). While earlier research on these gender-related language features was mainly intuition-based, recently, corpus-linguistic

research has confirmed many of these findings. However, corpus-based research has almost exclusively focused on inner-circle varieties of English, such as British English (e.g. Baker 2014), American English (e.g. Romaine 2001) or New Zealand English (e.g. Holmes 2001). Corpus-based research investigating gender-preferential language features in outer-circle varieties has only rarely been undertaken (one laudable exception being Fuchs and Gut (2012), however, who found that women generally use more intensifiers than men in ESL varieties). In the present paper, I would thus like to test if gender-related speaking styles also figure to the same extent in an institutionalized second language variety of English that has not attracted too much attention so far, i.e. Jamaican English. The overall aims of the present paper are thus (1) to test if gender-specific language variation also figures in outer circle varieties of English, and (2) to test if this (hypothesized) effect is stable when other sociolinguistic variables are also taken into consideration.

The database for this study is spoken component of the Jamaican component of the *International Corpus of English* (ICE-JA), which includes detailed speaker profiles with a variety of meta-data on the speakers. In order to test linguistic gender-related stereotypes in ICE-JA, I would like to present a study based on 1) hedges (e.g. *I guess*, *I think*), minimal responses (e.g. *mhm*, *yeah*) 3) stereotypically (fe)male particles (e.g. *shit* vs. *oh dear*) and 4) stereotypically (fe)male adjectives (e.g. *lovely* vs. *cool*). I annotated each occurrence for the speaker's GENDER, AGE and EDUCATIONAL LEVEL, as well as for the communicative context in which it was uttered (i.e. same-sex vs. mixed-sex conversations). Methodologically, I apply multifactorial regression analysis (cf. Gries 2013) to the data in order to test if any (combinations of) the available sociolinguistic variables has an effect on the speakers' use of stereotypically gender-related linguistic features in the outer circle. While the results of this analysis largely confirm previous findings on ENL speakers (e.g. concerning a higher number of hedges uttered by females), there are also some deviances from previous findings on ENL data (e.g. on stereotypically gender-related uses of certain adjectives such as *pretty* or *horrible*). Finally, I will discuss these findings with regard to to their implications on ESL gender research in general as well as on the analysis of spoken ESL data in particular.

References

Baker, P. (2014). Using Corpora to Analyze Gender. London/New York: Bloomsbury.

Fuchs, R. & Gut U. (2012). Do women use more intensifiers than men? Investigating gender and age-specific language use with the International Corpus of English. Paper presented at the 33rd conference of the International Computer Archive of Modern and Medievel English (ICAME 2012), University of Leuven.

Gries, S. Th. (2013). Statistics for Linguistics with R: A Practical Introduction. Berlin: Mouton de Gruyter.

Holmes, J. (2001). A corpus based view of gender in New Zealand English. In M. Hellinger & H. Bussman (Eds.), Gender Across Languages. The Linguistic Representation of Women and Men. Vol 1. Amsterdam/Philadelphia: John Benjamins, 115–136.

Holmes, J. (2013). An Introduction to Sociolinguistics. London/New York: Routledge.
Lakoff, R. T. (2004). Language and Woman's Place: Text and Commentaries. Rev. and expanded ed. New York: Oxford University Press.
Romaine, S. (2001). A corpus-based view of gender in British and American English. In M. Hellinger & H. Bußmann (Eds.), Gender Across Languages: The linguistic representation of women and men. Amsterdam/Philadelphia: John Benjamins, 153–175.
Tannen, D. (1994). Gender and Discourse. London: Oxford University Press.

• • •

Beke Hansen
University of Kiel
hansen@anglistik.uni-kiel.de

Gender paradox – Late female lead in the restructuring of deontic modality in Indian English

Sociolinguistic research on deontic modality is generally rare in first-language varieties of English (cf. Tagliamonte & Smith 2006:369), but it is non-existent in second-language varieties of English. In this talk, I will address this research gap by analysing the complex interaction between genderlectal variation, age-related variation and register variation in the choice between the semi-modal verb *have to* and the modal verb *must* in British English and Indian English. My study shows that text type differences play the most important part in the selection of *have to* over *must* in both varieties. *Have to* is more often used in more informal text types, whereas *must* is more often used in more formal text types (cf. also Biber 2004:192). This distribution is rather unsurprising, given that semi-modal verbs are common in spoken language. What is more surprising is the fact that text type emerges as the only strong predictor of variation in deontic modality in British English, while the social factors age and gender emerge as further predictors in Indian English. This indicates that the rise of *have to* has reached a saturation point in British English (cf. Mair 2015:136). Age is not a strong predictor of variation (any more), suggesting that the change in progress has come to a halt. Furthermore, gender differences have also disappeared as the change has become established (cf. Labov 2001:309). In Indian English, by contrast, social factors turn out to be strong predictors of variation in deontic modality and indicate that deontic modality is involved in ongoing change. Speakers in the younger age groups use *have to* more often than *must* compared to speakers in the older age groups, and female speakers lead the change towards innovative *have to* in the younger age groups (cf. also Lange 2012:190 for female speakers of Indian English as innovators). However, the gender pattern is reversed in the older age groups, with

female speakers showing a higher proportion of conservative *must* than male speakers. While the distribution according to gender in the youngest age group reflects the gender pattern that is most likely in frequency shifts in deontic modality (as a change from below), the conservative behaviour of women in the older age groups is unexpected. This pattern suggests a delayed onset of female lead in this linguistic change. It is open to question whether this pattern can be generalised to other linguistic variables and whether it can be linked to differences and changes in the roles of women. Clearly, further sociolinguistic studies across ENL, ESL and EFL varieties are necessary to understand the complex interaction between differing gender roles and genderlectal variation.

References

Biber, D. (2004). Modal use across registers and time. In A. Curzan & K. Emmons (Eds.), Studies in the history of the English language: Volume II: Unfolding conversations. Berlin: Mouton de Gruyter, 189-216.

Labov, W. (2001). Principles of linguistic change: Volume II: Social factors. Malden, Mass.: Blackwell.

Lange, C. (2012). The syntax of spoken Indian English. Amsterdam: John Benjamins.

Mair, C. (2015). Cross-variety diachronic drifts and ephemeral regional contrasts: An analysis of modality in the extended Brown family of corpora and what it can tell us about the New Englishes. In P. Collins (Ed.), Grammatical change in English world-wide. Amsterdam: John Benjamins, 119-146.

Tagliamonte, S. A. & Smith J. (2006). Layering, competition and a twist of fate: Deontic modality in dialects of English. Diachronica, 23 (2), 341-380.

• • •

Claudia Lange
Dresden University of Technology
Claudia.Lange@tu-dresden.de

Sven Leuckert
Dresden University of Technology
sven.leuckert@gmx.de

Indian English tag questions revisited: the role of gender and style in social interaction

Over the years, the impetus for studying tag questions has changed: ever since Lakoff (1975), gender-specific use of tag questions has been investigated by feminist linguists (e.g. Holmes 1995). Variety-specific differences in usage have also been noted, first in contrastive studies focusing on the two prominent Inner Circle varieties British and American English (e.g. Algeo 1988, 2006: 293-303, Tottie & Hoffmann 2006) and then extending the range to Outer Circle varieties (e.g. Columbus 2009). One specific realization of tag questions, namely the "invariant non-concord tag" such as *innit*? in colloquial British English, *eh*? in Canadian English or *no/na* in Indian English is listed as feature no. 165 in the *Mouton World Atlas of Variation* (Kortmann & Lunkenheimer 2012) and found to be ubiquitous in Asian Englishes with an attestation rate of 100% (Mesthrie 2012: 795) and very widespread in L2 Englishes with an attestation rate of 83% (Lunkenheimer 2012: 850). Hovever, corpus-based studies on Asian Englishes have only focussed on subsets of tag questions: Parviainen (2016) studied the distribution of invariant *isn't it*, which is a minority option within the Indian English tag repertoire, and Takahashi (2014) only considered canonical tag questions, which are even rarer in Indian English (cf Lange 2012: 205).

The present paper aims at a much finer level of granularity in considering the repertoire of tag questions in Indian English, highlighting the role of gender in social interactions. To this end, both a quantitative and a qualitative approach are chosen. Firstly, the data derived from the ICE-India conversation files and presented in Lange (2012: 195-234) will be revisited with a multiple regression analysis to determine the interplay of internal and external context factors for the occurrence of specific tag questions types. Secondly, subsets of the ICE-India conversations which display both higher and lower frequencies of tag questions will be analyzed in order to establish the whole range of linguistic forms which share the epistemic and affective functions ascribed to tags. Special attention will be given to patterns of social interaction and their linguistic correlates in all-female, all-male and mixed-group conversations.

With this combination of approaches, this paper attempts to contribute to the discussion initiated by Labov (2015), calling for a re-appraisal of our established expectations about gender as a sociolinguistic variable in "the wider range of social relations" (2015: 21) that may come to the fore in non-Western societies.

References

Algeo, J. (1988). The Tag Question in British English: It‘s Different, I‘n‘it. English World-Wide 9 (2), 171–191.

Algeo, J. (2006). British or American English? A Handbook of Word and Grammar Patterns. Cambridge: Cambridge University Press.

Cheshire, J. (2004). Sex and Gender in Variationist Research. In J. K. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), The Handbook of Language Variation and Change. Malden, MA: Blackwell, 423-443.

Columbus, G. (2009). A corpus-based analysis of invariant tags in five varieties of English. Language and Computers 69 (1), 401–414.

Holmes, J. (1995). Women, Men and Politeness. London: Longman.

Kortmann, B. & Lunkenheimer, K. (Eds.) (2012). The Mouton World Atlas of Variation in English. Berlin: de Gruyter Mouton.

Labov, W. (2015). The discovery of the unexpected. Asia-Pacific Language Variation 1 (1), 7-22.

Lakoff, R. (1975). Language and Woman’s Place. Harper & Row.

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

Lunkenheimer, K. (2012). Typological Profile: L2 Varieties. In Kortmann & Lunkenheimer (eds.), 845-872.

Mesthrie, R. (2012). Regional Profile: Asia. In Kortmann & Lunkenheimer (eds.), 785-805.

Moore, E. & Podesva, R. (2009). Style, indexicality, and the social meaning of tag questions. Language in Society 38, 447–485.

Parviainen, H. (2016). The invariant tag isn‘t it in Asian Englishes. World Englishes 35 (1), 98-117.

Takahashi, M. (2014). A comparative study of tag questions in four Asian Englishes from a corpus-based approach. Asian Englishes 16 (2), 101-124.

Tottie, G. & Hoffmann, S. (2006). Tag questions in British and American English. Journal of English Linguistics 34 (4), 283–311.

Valentine, T. (1991). Getting the message across: Discourse markers in Indian English. World Englishes 10 (3), 325–334.

• • •

Lucía Loureiro-Porto
University of the Balearic Islands
lucia.loureiro@uib.es

Genderlectal differences between democratic and colloquial uses of the language in Outer Circle varieties?

Among the Asian varieties of English, IndEng is usually considered more reluctant to undergo the recent linguistic innovations typically found in inner-circle varieties, while HKEng tends to occupy positions closer to the other end of the Asian scale (e.g. Xiao 2009, Collins 2013). However, no in-depth analysis has been conducted as for the differences between male and female speech in these two varieties so as to check whether the same current changes are more salient in one of the genders (presumably in women, as suggested by Labov 2001). This paper will address the issues of colloquialization and democratization, which are said to be responsible for some of the recent changes in inner-circle varieties of English. Thus, colloquialization or "a shift to a more speechlike style" (Leech et al. 2009: 239; Farrelly & Seoane 2012: 393) is said to be at the back of changes such as an increase in the use of the future marker *be going to* (Mair 1997), a decline in *wh-* relativizers, an increase in the frequency of *no* negation to the detriment of *not* negation, an increase of *let's* (Leech et al. 2009) and an increase in the use of contractions (Collins 2013), among others. Democratization, or "the removal of inequalities and asymmetries in the discursive and linguistic rights, obligations and prestige of groups of people" (Fairclough 1992: 201) is claimed to be at the roots of an increasing use of non-sexist language, including the replacement of *Mrs*/*Miss* with *Ms*, the use of neutral professional terms (e.g. *fire-fighter* instead of *fireman*), the use of gender-neutral or inclusive third person singular pronouns (singular *they*, or coordinate *he or she*, rather than generic *he*), a decline in the use of deontic modal *must* and a parallel increase of semi-modals *have to, need to*, and also a decline in the use of titular nouns such as *Mr, Dr* (Leech et al 2009: 259). Given that some of these markers have been shown to exhibit higher values in HKEng than in IndEng (cf. Suárez-Gómez 2014 on relativizers, Collins 2013 on contractions, Loureiro-Porto 2016 on the replacement of modal *must* with the semi-modals *have (got) to* and *need(to)*), this paper will explore whether intra-variety differences regarding colloquialization and democratization can be also observed at the genderlectal level, with data drawn from the private dialogue sections of ICE-IND and ICE-HK (S1A). The markers of colloquialization and democratization selected for this study are: (i) future marker *will* vs. *be going to*; (ii) *no* negation vs. *not* negation, (iii) contractions (iv) use of gender-neutral pronouns (*they* and *he or she*) vs. generic *he*, (v) use of gender neutral professional terms and (vi) modal verbs of obligation (*must* vs. semi-modals). An analysis of almost 7,000 cases of THEY and HE shows that HK speakers prefer singular THEY as an epicene pronoun, while IND speakers choose generic HE. Nevertheless, no significant differences are observed between males and females in any variety. An in-depth analysis of the other features will shed more light on this issue.

References

Collins, P. (2013). Grammatical colloquialism and the English quasi-modals: a comparative study. In J. I. Marín-Arrese, M. Carretero, J. Arús Hita & J. van der Auwera (Eds.), English Modality. Core, Periphery and Evidentiality (TiEL 81). Berlin & New York: Mouton de Gruyter, 155-169.

Fairclough, N. (1992). Discourse and Social Change. Cambridge: Polity Press.

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

Labov, W. (2001). Principles of Linguistic Change, vol II: Social Factors. Oxford: Blackwell.

Leech, G., M. Hundt, C. Mair & N. Smith. (2009). Change in Contemporary English: A Grammatical Study. Cambridge: CUP.

Loureiro-Porto, L. (2016). (Semi-)modals of necessity in Hong Kong and Indian Englishes. In E. Seoane & C. Suárez-Gómez (Eds.), World Englishes. New Theoretical and Methodological Considerations (Varieties of English Around the World, G57). Amsterdam: John Benjamins, 143-172.

Mair, C. (1997). The spread of the going-to-future in written English: A corpus-based investigation into language change in progress. In R. Hickey & S. Puppel (Eds.), Language History and Linguistic Modelling. A Festschrift for Jacek Fisiak. Berlin: Mouton de Gruyter, 1537-1543.

Suárez-Gómez, C. (2014). Relative clauses in Asian Englishes. Journal of English Linguistics 42 (3), 245-268.

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

• • •

Christoph Wolk
University of Giessen
christoph.b.wolk@anglistik.uni-giessen.de

Gender-based variation in part-of-speech patterns across the inner and outer circle

It is an often-reported finding that male and female speakers differ with regard to the distribution of parts of speech. For example, female speakers prefer the use of personal pronouns over nouns in a wide variety of studies in British English conversations (Rayson, Leech & Hodges 1997), in British dialectal oral history interviews (Wolk 2014) and in social media (Bamman, Eisenstein & Schnoebelen 2014). This study aims to add a cross-varietal perspective to this body of research. The central research

questions are, first, whether such gender differences also hold across different varieties of English and, if so, what precisely the relevant patterns are, second, the degree to which individual varieties follow or diverge from the global trend, and finally the relative magnitudes of gender- and variety-based differences.

To this end, I tap five ICE components spanning both inner and the outer circle varieties that have both speaker-level metadata and CLAWS7-based part of speech tagging, namely ICE-GB, ICE-Ireland, ICE-India, ICE-Singapore and ICE-Jamaica. On the methodological plane, I employ permutation-based analysis, a non-parametric approach that makes no distributional assumptions and takes individual speakers' behavior into account instead of simply averaging over them, and thereby alleviates issues resulting from dispersion. More specifically, I use a method first proposed by Nerbonne & Wiersma (2006) for comparing two groups and its extension by Wolk (2014) to allow simultaneous evaluation of an arbitrary number of groups. In this approach, the observed corpus frequencies are compared against a large number of simulated corpora generated by randomly resampling speakers from the original data. This allows the researcher to untangle patterns driven solely by individual speakers' preferences from those widely shared across the group. Both the distribution of individual parts of speech and that of bigrams, i. e. sequences of two parts of speech, will be considered.

The results largely confirm the patterns observed on other data sets; for example, female speakers generally show higher frequency than male speakers for personal pronouns, most forms of primary verbs, the negator *not*, and interjections, while male speakers prefer parts of speech associated with a nominal style, including articles, nouns, adjectives, prepositions, and numbers, as well as finite lexical verbs. Individual varieties may, however, differ in how strong these differences are; Indian English, in particular, shows a weakened differentiation for many of them, although the patterns remain significant within this variety. For unigrams, gender differentiation seems to be stronger than that based on variety, but as one moves towards larger sequences, variety-based frequency patterns begin to dominate gender-based ones, in particular for outer circle varieties.

References

Bamman, D., Eisenstein, J. & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. Journal of Sociolinguistics 18 (2), 135–160.

Rayson, P., Leech, G. N. & Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. International Journal of Corpus Linguistics 2 (1), 133–152.

Nerbonne, J. & Wiersma, W. (2006). A Measure of Aggregate Syntactic Distance. In J. Nerbonne & B Hinrichs (eds.), Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006, 82–90.

Wolk, C. (2014). Integrating Aggregational and Probabilistic Approaches to Language Variation. Freiburg: University of Freiburg.

• • •

Workshop 2

Corpora et comparatio linguarum: Textual and contextual perspectives

Workshop convenors:
Hilde Hasselgård & Signe Oksefjell Ebeling (University of Oslo)

Karin Aijmer
University of Gothenburg
karin.aijmer@eng.gu.se

The Swedish modal auxiliary *ska/skall* and English *shall*

The Swedish modal auxiliary *ska/skall* and the English *shall* are cognates. This raises the question to what extent they are used in the same way in the two languages. On the one hand, *shall* and *skall* are correspondences in non-fiction (90% of the examples). However inn fiction the meanings of the two auxiliaries diverge (only 10% of the examples are intertranslatable). A more frequent English correspondence of *ska* is *should* as in the following example:

> En gammal fin sherry ska serveras direkt ur buteljen.
> A fine old sherry should be served straight out of the bottle.

However, in other contexts *ska* is better translated by *will:*

> Jag ska ta mig an det
> I'll see to that

Ska/skall has several different functions in various verbal systems such as tense and modality (Teleman 1999: 312). The analysis raises a number of questions. Are *ska* and *skall* used in the same way? Is *ska* (or *skall*) primarily modal or temporal and what is the relation between the different meanings? In what way does *ska/skall* refer to a future event? Hilpert (2008) has suggested that the contexts in which *ska/skall* occurs (eg the semantic type of verb they collocate with) can help us to better understand how they are used and what meanings they convey.

Another possibility is to use contrastive or translational corpus data to test hypotheses about the meanings that *ska/skall* receives in different contexts and their frequencies. We are lucky to be able study the use of *ska/skall* on the basis of the

translational correspondences found in the English-Swedish Parallel Corpus. The present study sets out to investigate the translations of *ska* and *skall* into English with Swedish either as the source language or the target language. A preliminary look at the corpus (from Swedish into English only) shows that a large number of different correspondences are represented in the corpus (*Skall/ska* is for instance realized in English as a modal auxiliary (*shall, should, must, can*), a semi-auxiliary (*be going to, be about to, have to, be supposed to*), a clause (*I want to*, *I want you to*).

Several different meanings can be distinguished on the basis of translations. *Ska/skall* refers to future time in different ways as reflected in the translations into English (*will, be going to, be about to, be due to*, present tense).

When it has modal meaning *ska/skall* expresses evidential rather than epistemic) with reference to what people say modality (*be said to, be supposed to, they do say*. However, its modal meaning can also be deontic (*must, have to, have got to, should, need*). As a deontic modal auxiliary it is used in speech acts such as requests and offers. In the example below the meaning of *ska* has been rendered by an imperative:

Nu ska ni få höra, gubbar!
Now just you listen to me, you men.

A polite offer is exemplified by:

Ska faster inte ha en liten sherry i alla fall?
Would n't you like a little sherry after all, Auntie?

In the conclusion I will briefly discuss the differences between the English *shall* and its Swedish cognate *ska/skall* in terms of grammaticalization.

References

Hilpert, M. (2008). Germanic future constructions. A usage-based approach to language change. Amsterdam/Philadelphia: John Benjamins.

Teleman, U., Hellberg, S., Andersson, E. (1999). Svenska Akadeimiens grammatik. Stockhom Norstedts.

• • •

Karin Axelsson
University of Gothenburg
karin.axelsson@sprak.gu.se

Questions and responses in fiction texts in English and Swedish

Questions and their responses have been studied in many ways, a recent contribution being a volume edited by de Ruiter (2012). Such studies have usually been made on real-life conversation or invented examples, whereas questions in writing have received fairly little attention. Biber *et al.* (1999) report that questions are rare in newspaper language and academic prose, whereas there are many more questions in fiction, although there are less than a third compared to conversation. Biber *et al.* state that "the presence of dialogue accounts for the relatively high frequency in fiction" (1999:211) compared to other writing, but without giving any separate frequencies for questions in fiction dialogue. Proportionally, they find that *wh*-questions are more common in fiction overall than in conversation, whereas tag questions are less common. Indeed, Axelsson (2011) shows that tag questions are used about three times less often in fiction dialogue than in everyday real-life conversation. The aim of the present project is to investigate the distribution of question types in English fiction dialogue as well as outside the dialogue, and how these questions are responded to, in particular *yes/no*-questions. Enfield and Sidnell (2015) discuss two main strategies in such answers – interjection and repetition – predicting that "in all languages interjection confirmations tend to accept the terms of the question to which they respond whereas repeat confirmations are more assertive" (2015:11) (cf. also e.g. Hakulinen (2001) for Finnish and Bolden (2016) for Russian). As repetition (e.g. *I do/don't*) without an interjection/response word (e.g. *Yes/No*) seems less idiomatic in Swedish than in English, it is interesting to make a cross-linguistic study using a parallel corpus with these two languages, viz. the *English-Swedish Parallel Corpus* (ESPC). The study has started with the analysis of all examples with question marks in the English fiction originals (2,094 instances) together with some context, similar to the search Wikberg (1996) performed in the *English-Norwegian Parallel Corpus*. By conducting the initial search on questions, all types of answers will be included, also repetitions and other answers without a response word. Swedish original fiction texts will then be studied and compared to English original fiction texts. Finally, translations between the two languages will be investigated. The project also involves the mark-up of direct speech in ESPC, so that frequency calculations can be made on fiction dialogue separately as well as on text parts outside the dialogue.

References

Axelsson, K. (2011). Tag Questions in Fiction Dialogue. Ph.D., University of Gothenburg, Göteborg. URL: http://hdl.handle.net/2077/24047.

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

Bolden, G. B. (2016). A simple da?: affirming responses to polar questions in Russian conversation. Journal of Pragmatics 100, 40–58.
de Ruiter, J. P. (ed.) (2012). Questions: Formal, Functional and Interactional Perspectives. Cambridge: Cambridge University Press.
Enfield, N. J. & Sidnell, J. (2015). Language structure and social agency: confirming polar questions in conversation. Linguistics Vanguard 1 (1), 131–143.
Hakulinen, A. (2001). Minimal and non-minimal answers to yes-no questions. Pragmatics 11 (1), 1–15.
Wikberg, K. (1996). Questions in English and Norwegian: Evidence from the English-Norwegian Parallel Corpus. In C. E. Percy, C. F. Meyer & I. Lancashire (Eds.), Synchronic Corpus Linguistics: Papers from the Sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16). Amsterdam: Rodopi, 17–28.

• • •

Marcus Callies
University of Bremen
callies@uni-bremen.de

Magnus Levin
Linnaeus University
magnus.levin@lnu.se

A comparative corpus study of dislocation structures in live football commentary

This paper presents a comparative study of left and right dislocation as a functionally-motivated, register-specific feature of live sports commentary. Our data consist of comparable corpora of transcripts of the English, German and Swedish live TV commentaries of the 2014 men's football World Cup final. Dislocation serves information-structure purposes and involves a definite NP occurring in a peripheral position with a co-referential pronoun in the core of the clause (Biber et al. 1999: 956). The definite NP is found either left or right of the core clause as exemplified in (1a)–(1c) for the languages under study:

(1a) English
Well, they had a rocky few minutes, Germany, but they've seem to have gotten their rhythm back at the moment.

(1b) German
Er ist gut drauf, er ist gefährlich gut drauf, Lionell Messi.
'He's well on the ball, he's extremely well on the ball, Lionell Messi'

(1c) Swedish
Han gör det bra Müller som täcker undan Zabaleta i det där läget.
'He does it well, Müller, who shields the ball from Zabaleta in that situation'

The paper compares the (con-)textual functions of dislocation across the three Germanic languages. While left and right dislocation have been studied in some detail in all three languages individually (e.g. Geluykens 1987, 1992 and Tizón-Couto 2012 for English; Frey 2004 and Averintseva-Klisch 2008 for German; Teleman et al. 1999: IV: 440–9 for Swedish), it seems that a) a comparative study is missing and b) its pervasive use and specific functions in sports commentary have been largely overlooked (see Jürgens 2009: 169–170 for an exception with regard to dislocation in German). The paper argues that dislocation is highly frequent and has specific discourse functions in live TV football commentary. Similar to spontaneous conversation, left dislocation is largely used for topicalisation, while right dislocation functions as a repair mechanism to resolve referential ambiguity, especially when caused by the non-alignment of speech and image that is pervasive in live TV commentary of sports events.

References

Averintseva-Klisch, M. (2008). German right dislocation and afterthought in discourse. In A. Benz & P. Kühnlein (Eds.), Constraints in Discourse. Amsterdam: Benjamins, 225–247.

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

Frey, W. (2004). Notes on the syntax and pragmatics of the German left dislocation. In H. Lohnstein & S. Trissler (Eds.), The Syntax and Semantics of the Left Periphery. Berlin: Mouton de Gruyter, 203–233.

Geluykens, R. (1987). Tails (right-dislocations) as a repair mechanism in English conversation. In J. Nuyts & G. de Schutter (Eds.), Getting One's Words Into Line: On Word Order and Functional Grammar. Dordrecht: Fortis, 119–129.

Geluykens, R. (1992). From Discourse Process to Grammatical Construction: On left-dislocation in English. Amsterdam: Benjamins.

Jürgens, F. (2009). Syntaktische Formen bei der Fußballberichterstattung. In A. Burkhardt & P. Schlobinski (Eds.), Flickflack, Foul und Tsukuhara. Der Sport und seine Sprache. Mannheim: Dudenverlag, 160–174.

Teleman, U., Andersson E. & Hellberg, S. (1999). Svenska Akademiens Grammatik. Stockholm: Norstedts.

Tizón-Couto, D. (2012). Left Dislocation in English. A functional-discoursal approach. Frankfurt/Main et al.: Peter Lang.

• • •

Libuše Dušková
Charles University, Prague
libuse.duskova@ff.cuni.cz

Parallel corpora and contrastive linguistics: where to look for pitfalls

The paper addresses some of the problems involved in contrastive studies between English and Czech based on original texts and their translations. Even though the relevance of parallel texts to contrastive research is unquestionable, insofar as they provide two interlinguistically synonymous versions of an identical content, they also involve some pitfalls. One of them, viz. the influence of the original on the choice of a translation counterpart, specifically of English non-finite constructions on their rendering in Czech, was pointed out as early as the beginnings of English-Czech contrastive studies based on printed parallel texts excerpted manually (Vachek 1955; Hladký 1961).

Of the problems stemming from the differences between English and Czech the present paper largely focuses on those arising from the different hierarchy of the respective word order principles, the primary one being, respectively, grammatical function in English, and functional sentence perspective (information structure) in Czech (Firbas 1959, 1964; Dušková 2012, 2015a). Considering that the factors determining the FSP structure include, in addition to word order, context, semantics and intonation (Firbas 1992), and that the realization forms of the carriers of the FSP functions lack a distinctive form, studies in this field largely have to resort to manual search of digitalized parallel texts. Automatic search has so far been restricted to fixed syntactic structures such as verb complementation (Brůhová 2014), copular predicates (Malá 2014; Dušková 2012), the cleft sentence (Dušková 2015b; Kudrnová, forthcoming) and tough movement (Popelíková 2015). Of other points, attention is paid to the choices made by the target language where both languages have parallel structures. This point is considered in the broader context of the general feature of translation counterparts, viz. the chosen equivalent is hardly ever the only one that can be used, especially at the higher language levels, which raises the question of the reason for the particular choice.

The discussion starts from a general consideration of the different types of translation counterparts with respect to their relevance to the point under study. This is assumed to be connected with the respective language level (morphological; syntactic – phrasal, clausal; hypersyntactic) and the text sort. Another general point of translation counterparts taken into consideration is their semantic and informational adequacy, with failures in this respect representing the pitfall areas of this methodology.

The material part of the study is based on examples excerpted manually from the InterCorp.

References

Brůhová, G. (2014). Ditransitive complementation from the FSP point of view. Prague Studies in English XXVI, AUC Philologica 3 (2013), 61-83.

Dušková, L. (2012). Vilém Mathesius and contrastive studies, and beyond. In M. Malá & P. Šaldová (Eds), A Centenary of English Studies at Charles University: from Mathesius to present-day linguistics, 21-48. Charles University in Prague,

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

Dušková, L. (2015b). Textual roles of two forms of rhematic subjects: initial rhematic subjects vs. subject rhematized by it-clefts. Linguistica Pragensia 25/(1), 49-66.

Firbas, J. (1959). Thoughts on the communicative function of the verb in English, German, and Czech. Brno Studies in English 1, 39-68 (reprinted in Collected Works of Jan Firbas, 1, 126-56. Masaryk University Press, Brno 2010).

Firbas, J. (1964). From comparative word order studies. Thoughts on V. Mathesius's conception of the word-order system in English compared with that in Czech. Brno Studies in English 4, 111-128; reprinted in Collected Works of Jan Firbas, 1, 239-258, Masaryk University, Brno, 2010.

Firbas, J. (1992). Functional Sentence Perspective in Written and Spoken Communication. Cambridge: Cambridge University Press.

Hladký, J. (1961). Remarks on complex condensation phenomena in some English and Czech contexts. Brno Studies in English 3, 105-118.

Kudrnová, A. (forthcoming). English it-clefts with focused subject from an FSP perspective and their Czech translation equivalents. Prague Studies in English XVIII.

Malá, M. (2014) English Copular Verbs. A contrastive corpus-supported view. Charles University in Prague, Faculty of Arts, Trivium 5.

Popelíková, J. (2015). Tough-constructions and the issue of thematicity: a study of the word easy in 17th and 18th century English. Linguistica Pragensia 25/(1), 37-48.

Vachek, J. (1955). Some thoughts on the so-called complex condensation in Modern English. Sborník prací Filosofické fakulty brněnské univerzity A 3, 63-77.

InterCorp Books available at http://www.korpus.cz/intercorp

• • •

Thomas Egan
Inland Norway University of Applied Sciences
Thomas.Egan@inn.no

Contexts of failure

Predications of failure by means of the English verb *fail* may involve very different presuppositions. They may encode situations of disappointed effort on the part of the subject, as in (1), situations of neglected duty, again on the part of the subject, as in (2), and situations of thwarted speaker expectation, as in (3).

(1) It was a deep disappointment to him when the Scots *failed to gain* their objectives… (MH1)
(2) If a Member State persists in *failing to put into practice* the recommendations of the Council… (MAAS1)
(3) Incredibly, even now the German High Command *failed to perceive* either the invasion fleet or the significance of the massive Allied activity. (MH1)

These three examples, all taken from the English-Norwegian Parallel Corpus (ENPC: see Johansson 2007), may be situated on a cline of objectivity-subjectivity, in the sense of Traugott (2010: 33). The most objective sense is the 'Effort' sense, (1). The most subjective is the 'Expectation' sense, (3). The 'Duty' sense (2) is located towards the Effort end of the cline. In order to distinguish between these three senses, all of which are listed in the OED, we may have recourse to the co-text and/or context. One of the goals of the present paper is to investigate whether these three different senses differ in their Norwegian translation correspondences.

There is, moreover, a fourth sense of 'fail to', not distinguished as such in the OED, in which the predication is bleached of any notion of effort, duty or expectation. This sense is illustrated in (4).

(4) Another family of ceratoids *fails to develop* large nostrils …. (SJG1)

Egan (2010, 2016) refers to this as the 'Negation' sense and maintains that it displays many of the classical features of grammaticalisation, including what Boye and Harder (2009, 2012) call 'discursive backgrounding'. In this presentation I will pay particular attention to the translation correspondences of this negation sense of *fail.* I will also investigate its distribution across various contexts and text types.

The data for the study comprise all tokens of the verb *fail* in both original English texts and English translations in the ENPC. Both fictional and non-fictional texts will be examined with a view to determining whether the grammaticalised variant is more common in one or other text type. If the 'fail to' construction is undergoing grammaticalisation, one might hypothesise that it would be frequently translated by the default Norwegian negation marker 'ikke' (not). However, one would not expect 'fail to' to be employed to the same extent as a translation of Norwegian 'ikke', since the

default means of coding negation in English is by means of 'not'. One would therefore hypothesise that there would be more tokens of the 'fail to' construction in the English original texts than the English translations. In fact, in all three of the main syntactic roles of *fail*, intransitive, transitive with a nominal (frequently reflexive) object and with a *to*-infinitive, there are considerably more tokens in the English originals. All three construction types will be investigated and an explanation offered for these discrepancies.

References

Boye, K. & Harder, P. (2009). Evidentiality: Linguistic categories and grammaticalization. Functions of Language, 16 (1),9-43.

Boye, K. & Harder P. (2012). A usage-based theory of grammatical status and grammaticalization. Language, 88 (1),1-44.

Egan, T. (2010). The 'fail to' Construction in Late Modern and Present-day English. In U. Lenker, J. Huber & R. Mailhammer (Eds.) English Historical Linguistics 2008. Selected papers from the fifteenth International Conference on English Historical Linguistics. Volume 1: The history of English Verbal and Nominal constructions, Amsterdam: John Benjamins, 123-141.

Egan, T. (2016). The subjective and intersubjective uses of 'fail to' and 'not fail to'. In H. Cuyckens, L. Ghesquière & D. van Olmen (Eds.), Aspects of Grammaticalization: (Inter)subjectification and Pathways of Change, Berlin: De Gruyter, 167-196.

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

Traugott, E. C. (2010). (Inter)subjectivity and (inter)subjectification: A reassessment. In K. Davidse, L. Vandelotte & H. Cuyckens (Eds.), Subjectification, intersubjectification and grammaticalization, Berlin: De Gruyter, 29–71.

• • •

Lobke Ghesquière
Université de Mons
lobke.ghesquiere@umons.ac.be

Lieselotte Brems
Université de Liège
lbrems@ulg.ac.be

Time, preference and intensity: A contrastive study of *rather* (*than*) and *eerder* (*dan*)

This paper reports on a synchronic corpus study of the English adverbial marker *rather (than)* and its Dutch equivalent *eerder (dan)*, which can both express both preference or contrast (1a, 2a), and intensity (1b, 2b). In addition, Dutch *eerder (dan)* has temporal uses as in (3), which are no longer attested for English *rather (than)*. As such, it can be hypothesized that *rather (than)* has progressed further than *eerder (dan)* along the grammaticalization pathway leading from temporal to preferential meaning posited by a.o. Traugott & König (1991: 203–204). This pathway has been confirmed for English by a.o. Rissanen (2008) and for French *plutôt (que)* by Mokni (2008), but to date no cross-linguistic study or corpus study on Dutch *eerder (dan)* has been carried out. Based on synchronic data from comparable and parallel corpora, we will draw up typologies of the different uses of *rather (than)* and *eerder (dan),* comparing them qualitatively by characterizing them semantically and structurally, and quantify the results.

(1) a. Unfortunately, this adequacy was a reminder that his problem has not been his lack of style but rather his abundance of insincerity (Rissanen 2002: 357)
    b. Sachs understood that she was playing with him but he rather enjoyed the way she went about it. (OED, rather, 6b)
(2) a. Maar ik noem dat geen fatalisme, ik zie er eerder een vorm van verweer in. (DBNL)"
    'But I wouldn't call it fatalism. I'd rather see a kind of defence in it'
    b. En hoewel de Zweden van nature uit een eerder stijf en nauwgezet volk zijn kan men er dan ook in elke krantenkiosk, en gewoon tentoongesteld, tijdschriften zien met foto's die onze zedenmeesters de kolieken zouden doen krijgen. (DBNL)
    'And even though the Swedes are by nature a rather stiff and meticulous people one can find displayed in every newspaper stand magazines with pictures that would give our moralists the gripes'
(3) Waarom werd dat niet eerder aan de orde gesteld? (DBNL)
    'Why wasn't this matter brought up sooner?'

This study is based on both translated and original language which allows us to assess the degree of intertranslatability of the two constructions, their similarity or

difference in use and frequency in original and translated text as well as to come to a better understanding of the different language-internal uses of *rather (than)* and *eerder (dan)*. The monolingual English corpus used is the British books section of the Wordbanks Online corpus. For Dutch we queried a selection of 20th and 21st century texts of the *Digitale Bibliotheek voor de Nederlandse Letteren* [Digital Library of Dutch Literature] (DBNL) (2.3 million words). The translation data for this study are extracted from the bi-directional Dutch Parallel Corpus (DPC), a 10-million-word parallel corpus comprising texts in Dutch, English and French with Dutch as a pivotal language. Text types range from literary prose to nonfiction material, such as essays or newspaper, business, technical and policy texts.

The parameters taken into account for the data analysis include the syntactic realization of the elements under scrutiny (as adjunct, subjunct, conjunct or modifier; Quirk *et al.* 1985), their collocational patterning and their specific pragmatics. For the contrastive uses, a fine-grained typology will be set up of the precise textual relations expressed (contrast, reformulation, replacement, etc.; Quirk *et al.* 1985: 638–639). For the intensifying uses, the specific types of degree modification will be charted (e.g. upscaling vs. downscaling, Quirk *et al.* 1985; open vs. closed scale modification, Kennedy & McNally 2005).

Corpora

DBNL: *Digitale Bibliotheek voor de Nederlandse Letteren* [Digital Library of Dutch Literature]. Available online at http://www.dbnl.org.

DPC: Dutch Parallel Corpus. Available online at https://www.kuleuven-kulak.be/dpc/conc/.

WB: Wordbanks Online. Available online at https://wordbanks.harpercollins.co.uk/.

References

Kennedy, C. & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. Language, 81(2), 345–381.

Mokni, M. (2008). La grammaticalisation de l'adverbe plutôt et l'évolution du système grammatical. Linx, 59, 171–184.

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

Rissanen, M. (2008). From 'quickly' to 'fairly': On the history of ‚rather'. English Language and Linguistics, 12(2), 345–359.

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

• • •

Anne-Line Graedler
Inland Norway University of Applied Sciences
anneline.graedler@inn.no

Contrastive analysis of L2 students' cohesive choices in translation

Translation has traditionally been a component of advanced foreign language study programs, and is also about to become reestablished as an effective tool of L2 learning, partly based on arguments connected to the notion of the awareness of language in general, and of contrastive aspects in particular (Whyatt 2009; Cook 2010). Contrastive discourse analysis gives emphasis to both similarities and differences in discourse organization between different languages (cf. e.g. Taboada, Suárez & Álvarez 2013). It is also one of the main areas for both study and evaluation of translation (Baker 2011: 131ff; House 2009: 31-32), and furthermore, it is highly relevant for general assessment of language learners' textual competence. This progress report will focus on language learners' competence and awareness of text coherence from a cross-linguistic perspective, i.e. how grammatical cohesion in source texts is transferred or changed in target texts translated from L1 Norwegian into L2 English. The main goal is to analyze what types of grammatical cohesive devices are displayed in the translation source texts, and whether and in what way the target texts (translations) display equivalent (or non-equivalent) cohesive relations.

Cohesive texts are created in different ways, and the focus will here be grammatical cohesion based on structural content, such as the following categories (Halliday & Hasan 1976):

- reference, i.e. the relationship of identity between linguistic expressions;
- substitution and ellipsis, i.e. the replacement or omission of language items by grammatical structures; and
- conjunction, i.e. the use of different items to relate different parts of the text to each other.

The research data is the Norwegian-English Student Translation corpus (NEST), which consists of 32 Norwegian source texts and 348 English target texts. The texts were submitted by 141 advanced learners of English from different tertiary level institutions in Norway, participating in translation courses as part of regular English studies, not specific translation programs. The NEST corpus also contains metadata about the students' language and educational background. In addition to the corpus data, a number of commentary texts and essays exist where the students reflect upon and explain choices related to various aspects of the translation in the NEST texts; something that may be included in the analysis if relevant for the selected topics above.

In addition to a comparison between cohesive devices in English and Norwegian (cf. Faarlund, Lie & Vannebo 1997), the study will thus contribute to explore and discuss the role of contrastive discourse analysis in relation to translation assignments for advanced English language learners, and also their general competence and awareness of text coherence.

References

Baker, M. (2011). In Other Words: A course book on translation (2nd ed.). London & New York: Routledge.

Cook, G. (2010). Translation in language teaching: An argument for reassessment. Oxford: Oxford University Press.

Taboada, M., Suárez, S.D. & Álvarez, E.G. (Eds.) (2013). Contrastive Discourse Analysis. Functional and Corpus Perspectives. Sheffield: Equinox.

Faarlund, J.T, Lie, S. & Vannebo, K.I. (1997). Norsk referansegrammatikk. Oslo: Universitetsforlaget.

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

House, J. (2009). Translation. Oxford: Oxford University Press.

Whyatt, B. (2009). Translating as a way of improving language control in the mind of an L2 learner. In A. Witte et al. (Eds.), Translation in Second Language Learning and Teaching. Bern: Peter Lang, 181-202.

• • •

Hilde Hasselgård
University of Oslo
hilde.hasselgard@ilos.uio.no

Indefinite subjects in English and Norwegian

The present study investigates the frequency and use of indefinite noun phrases as subjects in English and Norwegian. Both languages are assumed to select subjects in agreement with information structure. Since the beginning of the sentence – the subject position – is associated with given information, definite noun phrase realization is expected (Prince 1992). Indefinite NPs are typically associated with new information in both languages, and hence disfavoured as subjects (Biber et al. 1999: 269). However, indefinite NP subjects are found in both languages, as shown by (1) and (2).

(1) *En tyv* er ikke voldsom, men stillferdig. (KF1)
A thief is not violent but quiet. (KF1T)
(2) *A rat* crept out of the hole behind the dresser … (GN1)
En rotte kom ut av hullet bak kommoden … (GN1T)

English and Norwegian indefinite NPs are formally and functionally rather similar: there are indefinite articles for the singular only (*a/an* and *en/ei/et* respectively), while indefinite plurals occur with no article or with certain types of quantifying determiners (e.g. cardinal numbers, some; cf. Lyons 1999: 33 ff). Preliminary analysis indicates that indefinite subjects are less frequent in Norwegian than in English. This

was expected on the basis of previous contrastive studies of thematic structure in the two languages (Hasselgård 2005).

The material consists of declarative main clauses from the fiction part of the English-Norwegian Parallel Corpus (ENPC). As the corpus is not parsed, subjects were retrieved using a combination of PoS tagging and positional criteria: i.e., searches were made for indefinite articles and indefinite plurals in first and second position in an s-unit. The concordances were sorted manually to exclude irrelevant hits.

The study first compares original texts in English and Norwegian asking the following research questions: How frequent are indefinite subjects in English and Norwegian? What are the lexicogrammatical features of such subjects (e.g. +/- modification) and their verbs (e.g. +/- transitive)? What are the contexts for indefinite subjects – and are they the same in both languages? For example, in both languages indefinite subjects occur in contexts where the subject has generic reference, as in (1), and in clauses denoting the existence or appearance of a specific subject referent, as in (2). Some of these may be characterized as bare presentatives (Ebeling 2000: 157), as illustrated by (3). Place adverbials are another frequent feature of clauses with indefinite subjects.

The second part of the study probes further into cross-linguistic differences in the use of indefinite subject NPs by exploring their translations. Given that indefinite subjects are comparatively rare in both languages, the translation principle of normalization might prompt translators to make changes to either the subject NP (as in (3)) or the whole clause, as in (4).

(3) *Cultured pearls* are in the vault. (DF1)
Kunstperlene ligger i velvet. (DF1T) [The cultured pearls lie in the vault.]

(4) *En gammel kvinne* tok imot oss, vennlig, men uten smil. (JW1) [An old woman received us…]
We were received *by an old lady*, in a friendly but unsmiling fashion. (JW1T)

Since indefinite subjects are less frequent in Norwegian than in English, translations into Norwegian are expected to involve a change of the subject NP more often than translations into English, especially if the NP has specific reference, as is the case in (3) and (4). Information structure and semantics, especially the notions of existence/appearance are also expected to play a role.

References

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). Longman Grammar of Spoken and Written English. London: Longman.

Ebeling, J. (2000). Presentative constructions in English and Norwegian. A corpus-based contrastive study. Doctoral thesis. Oslo: Faculty of Arts, University of Oslo.

Hasselgård, H. (2005). Theme in Norwegian. In K.L. Berge & E. Maagerø (eds), Semiotics from the North: Nordic Approaches to Systemic Functional Linguistics. Oslo: Novus, 35-48.

Lyons, C. (1999). Definiteness. Cambridge: Cambridge University Press.
Prince, E. F. (1992). The ZPG letter: Subjects, definiteness, and information-status. In W.C. Mann & S. A. Thompson (eds.) Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text. Amsterdam: John Benjamins Publishing Company, 295–326.

• • •

Mats Johansson
Centre for Languages and Literature, Lund University
mats.johansson@englund.lu.se

Lene Nordrum
Centre for Languages and Literature, Lund University
lene.nordrum@englund.lu.se

Swedish *orka* viewed through its English correspondences – insufficient energy or couldn't be bothered?

Swedish *orka* is interesting in a contrastive perspective since it lacks a straightforward equivalent in English. This paper considers instances of *orka* and its English correspondences in the *English-Swedish Parallel Corpus* – both in the direction *Swedish original* to *English translation* and in the direction *English original* to *Swedish translation*. Following the method 'seeing language through multilingual corpora' outlined by Johansson (2007), the aim of this paper is to shed light on the meaning and use of Swedish *orka*. By approaching the corpus correspondences from two different angles; focusing first on congruent correspondences and then on the semantic features in the context of such correspondences, we suggest that Swedish *orka* involves complex modality where the English correspondences often directly reflect one modal category by means of a congruent translation with a modal auxiliary or semi-auxiliary, but bring in other meaning categories by means of context. Specifically, we argue that the contextual patterns in our material underline *orka* as a so-called sufficiency verb, where the semantic concept *sufficiency* (Fortuin 2013) is a core semantic property. Sufficiency is an interesting concept cross-linguistically since it has been associated with modal notions such as the possibility of realizing contextually given goals. In this study, we consider particularly the distinction between sufficiency as a modal category and more central modal categories such as *ability*. We also note that *orka* predominantly occurs in negative polarity contexts, which might suggest that it is a negative polarity item with specific licensing conditions in the immediate linguistic environment. Apart from offering a specific semantic description of the Swedish verb *orka*, our study contributes to cross-linguistic studies of expressions of *sufficiency*

(Fortuin 2013) as well as to the study and possible definition of negative polarity items.

References

Fortuin, E. (2013). The construction of excess and sufficiency from a crosslinguistic perspective. Linguistic Typology 17, 31-88.

Johansson, S. (2007). Seeing through Multilingual Corpora. Amsterdam/Philadelphia: John Benjamins.

• • •

Ekaterina Lapshinova-Koltunski
Saarland University
e.lapshinova@mx.uni-saarland.de

Kerstin Kunz
Universität Heidelberg
kerstin.kunz@iued.uni-heidelberg.de

English vs. German from a Textual Perspective: Looking inside the Chain Intersection

This paper presents a corpus-based analysis on the interaction of coreference and lexical cohesion in English and German. Our previous results show that English prefers more lexical means for establishing cohesion, whereas German tends to use more coreference. On the one hand, this may be attributable to typological differences between the two languages: The German language system provides more linguistic means for expressing coreference, especially demonstrative ones, e.g. pronominal adverbs such as *damit* (thereby), *darüber* (thereover) that are not common in Present Day English. On the other hand, conceptual preferences might be at play, with German favouring to explicitly express topic continuity while English may tend towards topic progression.

Quite often, linguistic means of the two types of cohesion are combined, causing an intersection between chains of cohesion and coreference and, hence, an interaction between topics. An example of this kind of intersection is shown in (1), where we have a coreference chain with the elements *reservoirs — them — they* and a lexical chain with the elements *reservoirs — reservoirs — service reservoirs*. We focus on this intersection and analyse elements of lexical cohesion chains that occur inside of coreference chains. This will allow us to explore variation in semantic relations inside coreference chains, and will reveal which of the two languages contain more intersectional elements. It will complement our previous studies (Kunz et al. 2016) on topic

development in which we analyzed features of coreference chains and lexical chains (e.g. distance of elements in chains, chain length and number of chains) separately.

(1) *Well, in Edinburgh most of the water comes from <reservoirs> which are more towards the a lot of <them> are more towards the Borders, and then <they>'re actually quite old, I think. They first tried to sort out water in Edinburgh, and I think in about 1700... they slowly started to introduce laws to try to improve the public water supply... But now, a lot of the water comes from the hills on the outskirts of Edinburgh. And then it's brought into holding <reservoirs>, and then it's brought into the treatment works... And it's the company looks at different ways of helping water companies manage those assets better... if you've got a set of pipes and you've got a set of, say, <service reservoirs>, which is where you store the clean water, then if something happens in the system, how long have you got before the whole network runs out of water?*

We concentrate on the following questions: (1) Which language has a greater amount of intersectional cases? (2) How many antecedents are shared by coreference and lexical chains? (3) What are the most frequent semantic relations that occur within coreference chains in both languages?

We use a corpus annotated for both lexical cohesion (Martinez et al., 2016} and coreference chains (Lapshinova & Kunz, 2014), and filter out the chains that represent the cases of intersection. Our preliminary analyses show that the English texts in this dataset show more cases of intersection than the German ones (0.038 vs. 0.026). In terms of semantic relations, we observe more general ones (hyponyms) in English, whereas German prefers more specific ones. At the same time, repetitions (expressing identity) are the most common lexical members within coreference chains in both languages.

In our presentation, we will report on the results and our interpretation in terms of contrasts in topic development between English and German.

References

Kunz, K., Lapshinova-Koltunski, E., Martínez Martínez, J.M. (2016). Beyond Identity Coreference: Contrasting Indicators of Textual Coherence in English and German. In Proceedings of the workshop on Coreference Resolution Beyond OntoNotes co-located with NAACL-2016, San Diego, California, ACL.

Martínez Martínez, J.M., Lapshinova-Koltunski, E., Kunz, K.A. (2016). Annotation of lexical cohesion in English and German: Automatic and manual procedures. In Proceedings of the Conference on Natural Language Processing (Konferenz zur Verarbeitung natürlicher Sprache) – KONVENS-2016, Bochum, Germany, 165-176.

Lapshinova-Koltunski, E., Kunz, K. (2016). Annotating cohesion for multillingual analysis. In Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, Reykjavik, Iceland, LREC, 57-64.

• • •

Rosa Rabadán
University of León, Spain
rosa.rabadan@unileon.es

Veronica Colwell
University of León, Spain
veronica.colwell@unileon.es

Hugo Sanjurjo-González
University of León
hsang@unileon.es

Food in context: yumminess and yuckiness in Spanish-into-English restaurant menus

The explosion of tourism has given rise to the spread of new forms of enjoying foreign lands and cultures. One of them is the so-called culinary tourism. In (northern) Spain this steadily growing trend has revealed cross-cultural and cross-linguistic problems to the point of triggering new regulations for local restaurateurs. When communicating local cuisine, corpus-based contrast can help provide a way of responding to these challenges

This study focuses on Spanish restaurant menus and their translations into English as offered to guests and pays special attention to i) 'institutionalized' denominations of traditional dishes - metaphoric uses which are perfectly clear in context but tend to be opaque for the visitors (Falzett 2012); ii) temperature and quantity expressions (Ameka 2015), and iii) qualifying resources (Manca 2008; Diederich 2015).

The starting point of our study is twofold: i) a parallel corpus En-Es of 50 bilingual En-Es mid-range, independent, conventional restaurant menus, and ii) a custom-made, En-Es comparable corpus of recipes containing 135,912 words in the En subcorpus and 145,449 in the Es one. Both subcorpora are rhetorically and semantically annotated. This has been done using the ACTRES rhetorical tagger http://contraste2.unileon.es/web/en/tagger.html along with a tentative set of genre-and-domain -restricted semantic tags that conform to the UCREL USAS system http://ucrel.lancs.ac.uk/usas/ (Category F, F1 Terms relating to food and food preparation. E. g. Fare; Procedure; Ingredients, etc.)

The parallel corpus has revealed serious problems in communicating with international clients as the menus frequently fail to meet intelligibility expectations in the three aforementioned areas (Bubel & Spitz 2013).

The comparable corpus has been used as a source of additional information about cooking styles (e. g. *a la gallega*-Galician style: boiled food accompanied by olive oil, salt and paprika), which are frequently a source of context-bound metaphors. Cross-linguistic (dis)similarities in characterizing routines and woolliness with regards to amounts and temperatures, which tend to be associated with the culinary preparation and not openly encoded, can also be addressed using comparable data.

Corpus data are used to a) unveil functional equivalents whenever possible and b) finding positive- or at least neutral- ways of providing acceptable, reassuring descriptions of dishes. In addition, these cross-linguistic findings are being used to construct a controlled natural language of the drafter type (Hartley and Paris, 1996 & 1997; Power, Scott and Evans 1998; Kuhn 2014) that will help bridge the cross-linguistic and cross-cultural gap by offering more effective genre-and-domain restricted bi-textual options.

References

Ameka, F. K. (2015). "Hard sun, hot weather, skin pain": The cultural semantics of temperature expressions in Ewe and Likpe (West Africa). In M. Koptjevskaja-Tamm (ed), The Linguistics of Temperature, Amsterdam: Benjamins, 43–72.

Bubel, C. & A. Spitz. (2013). The way to intercultural learning is through the stomach – Genre-based writing in the EFL classroom. In C. Gerhardt, M. Frobenius & S. Ley (Eds.) Culinary Linguistics: The chef's special. Amsterdam: Benjamins, 157–188. DOI: 10.1075/clu.10

Diederich, C. (2015). Sensory Adjectives in the Discourse of Food. A frame-semantic approach to language and perception. Amsterdam: Benjamins. DOI: 10.1075/celcr.16

Falzett, T. (2012). "Bhio' tu dìreach ga ithe, bha e cho math = You would just eat it, it was so good". Music, Metaphor and Food for Thought on Scottish Gaelic Aesthetics. In A. Idström & E. Piirainen (Eds), Endangered Metaphors. Amsterdam: Benjamins, 315–338 DOI: 10.1075/clscc.2.15fal

Hartley, A. F. & Paris, C. (1996). Automatic text generation for software development and use. In H. L.Somers (ed), Terminology, LSP and Translation: Studies in language engineering in honour of Juan C. Sager. Amsterdam: Benjamins, 221-242.

Hartley, A. F. & Paris, C. (1997). Multilingual document production: from support for translating to support for authoring. Machine Translation, Special Issue on New Tools for Human Translators, 12 (1-2), 109-129.

Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. Computational Linguistics, 40 (1), 121-170. doi:10.1162/COLI_a_00168

Manca, E. (2008). From phraseology to culture: Qualifying adjectives in the language of tourism. In U. Römer & R. Schulze (Eds.), Patterns, meaningful units and specialized discourses. Special Issue of International Journal of Corpus Linguistics 13 (3), 368–385

Power, R., Scott, D., & Evans, R. (1998). What you see is what you meant: direct knowledge editing with natural language feedback. In H. Prade (Ed.), Proceedings of ECAI 98, 13th European Conference on Artificial Intelligence. NY: Wiley & Sons, 677-681.

• • •

Sylvi Rørvik
Inland Norway University of Applied Sciences
sylvi.rorvik@inn.no

Marte Monsen
nland Norway University of Applied Sciences
marte.monsen@inn.no

Marked themes in English and Norwegian academic texts in the field of didactics

At Inland Norway University of Applied Sciences, there is a master's program in the teaching and learning of language subjects during which the students, specializing in either English or Norwegian, have to write a report on a small empirical study of a topic related to didactics. Hence, the texts they produce are a ready-made comparable corpus of novice academic writing. The present study is a first attempt at providing insights about contrastive differences between Norwegian and English that may be used to inform novice writers about central textual features of expert academic texts within the field of didactics. To this end, a contrastive study was carried out of texts by expert writers of English and Norwegian (i.e. published academics), alongside a comparison of the above-mentioned student texts. It was decided to focus on the frequency and realization of marked themes (roughly defined as non-subjects in initial position, see further Halliday 2004: 64-105), as well as the meanings expressed by these, since previous research has presented conflicting evidence regarding the existence of contrastive differences in this area in other text types: for instance, in a comparison of fiction texts from the English-Norwegian Parallel Corpus, it was shown that Norwegian permits a higher frequency of non-subjects in initial position than does English (Hasselgård 1998, 2005). On the other hand, a study of argumentative newspaper texts found no contrastive difference as regards the frequency of marked themes (Rørvik 2013: 51-52).

The corpus consists of 11 published research articles in Norwegian and 11 in English, in addition to 11 student texts in L1 Norwegian and 11 in L2 English. The texts were divided into T-units (cf. Fries 1995: 318), and manually coded. Statistical calculations were carried out to compare the results for each corpus, by means of a one-way ANOVA with a Tukey post-hoc test.

The results show that there are no significant differences in the proportion or realization of marked themes between the Norwegian and English expert texts, nor when it comes to the distribution of meanings expressed by the marked themes. However, there are several areas where the two groups of novice writers differ from each other, for instance as regards the types of constructions they employ as marked themes: dependent clauses are more frequent in English than in Norwegian, while the opposite is true for prepositional phrases. Given that the expert texts do not exhibit the same differences, we conclude that the novice writers need advice in these areas in order to conform to the conventions of the text type and field.

References

Fries, P. H. (1995). Themes, development and texts. In R. Hasan & P. Fries (Eds.), On Subject and Theme. Amsterdam: John Benjamins, 317-359.

Halliday, M. A. K. (2004). An Introduction to Functional Grammar. 3rd edition, revised by C. M. I. M. Matthiessen. London: Arnold.

Hasselgård, H. (1998). Thematic structure in translation between English and Norwegian. In S. Johansson & S. Oksefjell (Eds.), Corpora and Cross-linguistic Research: Theory, Method, and Case Studies. Amsterdam: Rodopi, 145-167.

Hasselgård, H. (2005). Theme in Norwegian. In K. L. Berge & E. Maagerø (Eds.), Semiotics from the North. Nordic approaches to systemic functional linguistics. Oslo: Novus Press, 35-47.

Rørvik, S. (2013). Texture in learner language. Doctoral dissertation, University of Oslo.

• • •

Jenny Ström Herold
Linnaeus University
jenny.strom.herold@lnu.se

Magnus Levin
Linnaeus University
magnus.levin@lnu.se

Translating textual indeterminacy: English supplementive *ing*-clauses and their German and Swedish translations

Subordinatorless supplementive *ing*-clauses (as in *Hitler exploded, demanding examples*) are characterized by their "considerable indeterminacy as to the semantic relationship to be inferred" (Quirk et al. 1985: 1123; cf. also Biber et al. 1999: 820; Malá 2005) and may induce, for example, temporal, causal, conditional, concessive or circumstantial readings (Quirk et al. 1985: 1124; Kortmann 1991: 114–141). This puts a heavy interpretation load on translators, the translation task becoming more intricate with target languages lacking a similarly productive form. Thus, it is not unexpected that previous studies on the translation of *ing*-clauses into Swedish highlight the multitude of translation equivalents found in target texts (Lindquist 1989: 120–128; Blensenius 2006: 36).

This multi-target-language study provides insights into how different translators interpret and render the very same instance of an *ing*-clause. It draws on data from a new corpus being compiled by the authors. It comprises recently published English, German and Swedish non-fiction texts and their translations into the respective lan-

guages. In order to compare original English with translated English, the study also includes *ing*-clauses used as equivalents of German and Swedish source-text structures.

Preliminary findings indicate that the position of the *ing*-clause affects the distribution of translation equivalents. The rare sentence-initial *ing*-clauses more often show congruent, i.e. formally and semantically matching, translations, both target languages often opting for causal or temporal finite subclauses (e.g., *Feeding bread to the ducks, I noticed […]* rendered as *Als ich die Enten mit Brot fütterte* ('when I the ducks with bread fed') in German and similarly in Swedish *När jag matade änderna med bröd,* or a PP expressing manner (e.g., *Using this set up, Ellington's team could […]* rendered as *Mithilfe dieser Vorrichtung* ('by means of this device') in German and *Med detta upplägg* (*'with this set-up'*) in Swedish). The equivalents of sentence-final *ing*-clauses are more varied and non-congruent, ranging from infinitives to relative clauses, separate or coordinated main clauses, adverbial finite subclauses and prepositional phrases. Also, the translations more often differ semantically in this position. Overall, finite structures dominate in both target languages, which indicates that translators are producing more explicit structures. Moreover, our findings suggest that *ing*-clauses are much rarer in English target texts (in particular from Swedish originals), indicating a potential case of translationese.

References

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

Blensenius, K. (2006). Particip med andra ord. En korpusstudie av svenska motsvarigheter till engelska ing-satser. Göteborg: Göteborgs universitet, institutionen för svenska språket.

Kortmann, B. (1991). Free Adjuncts and Absolutes in English. London: Routledge.

Lindquist, H. (1989). English Adverbials in Translation: A Corpus Study of Swedish Renderings. Lund Studies in English 80. Lund: Lund University Press.

Malá, M. (2005). Semantic roles of adverbial participial clauses. Theory and Practice in English Studies 3: Proceedings from the Eighth Conference of British, American and Canadian Studies. Brno: Masarykova univerzita, 91–97.

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

• • •

Denisa Šebestová
Charles University
sebestovadenisa@gmail.com

Markéta Malá
Charles University
Marketa.Mala@ff.cuni.cz

The expression of interpersonal functions in Czech and in English: the Czech suffix *–pak* and its translation counterparts

The study explores the possibility to use translation counterparts as 'markers' (Malá 2013), or 'methodological anchors' (Gast 2015) of discourse functions, i.e. formal correlates of interpersonal and textual functions, which make it possible to detect these functions in the text, and compare their expression cross-linguistically.

We focus on Czech expressions containing the suffix *–pak* (such as, *copak* – '*what* + *pak*', *kdepak* – '*where* + *pak*'). The pronouns, adverbs, particles and interjections ending in *–pak* have been described as elements of 'the third syntactical plan' (Poldauf 1964), relating the content of an utterance to the individual and "his specific ability to perceive, judge and assess" (ibid.: 242). While the same communicative functions are expressed in Czech and in English, the means of expression as well as the extent to which the functions are explicitly marked are likely to differ in the two languages (example 1).

Jestlipak máte ještě tu tlustou knihu?
(Lit: Jestlipak (particle) you-have still that thick book?)
I wonder if you still have that thick book?

The study pursues two closely intertwined goals. First, the English translation counterparts of the Czech *–pak* expressions are identified, and their discourse functions described. Second, the English constructions are employed to further specify the pragmatic functions, style characteristics and preferred contexts of the Czech suffix *–pak*.

The suffix *–pak* is shown to be a polyfunctional indicator of communicative function (Grepl & Karlík 1998): the expressions ending in *–pak* have content/speaker-related functions (such as deliberative meaning, emotional evaluation, (im)possibility) as well as communication/addressee-oriented functions (appeal, establishing/maintaining contact) (Aijmer 2013; Šebestová & Malá 2016).

The English counterparts constitute a scale ranging from specific sentence types (e.g. negative rhetorical questions, exclamative sentences) via lexico-grammatical structures of varying degrees of fixedness (*I wonder if…* in ex. 1) to individual lexical markers of the discourse functions (e.g. intensifiers, lexical negators). In both languages, the indicators of the discourse function tend to be clause-initial: in English this applies, for instance, to conjunctions reinforcing the contact-maintaining func-

tion, negative idiomatic constructions (e.g. *not a hope*), or interrogative clause-opening expressions (e.g. *I wonder if…*, *Is it that…*).

The material was drawn from a Czech-English fiction subcorpus of the parallel translation corpus *InterCorp*. 560 Czech expressions comprising the suffix *–pak* and their English translation correspondences were analysed. Monolingual corpora were used to verify the findings based on the translation data.

References

Aijmer, K. (2013) Understanding Pragmatic Markers. A Variational Pragmatic Approach. Edinburgh: Edinburgh University Press.

Gast, V. (2015) On the use of translation corpora in contrastive linguistics. A case study of impersonalization in English and German. Languages in Contrast 15 (1), 4–33.

Grepl, M. & P. Karlík (1998) Skladba češtiny. Olomouc: Votobia.

Malá, M. (2013) Translation counterparts as markers of meaning. Languages in Contrast 13 (2), 170–192.

Poldauf, I. (1964) The Third Syntactical Plan. Travaux linguistiques de Prague 1, L´école de Prague aujourd´hui, 241-55.

Šebestová, D. & Malá, M. (2016) Anglické překladové protějšky českých vět s částicemi copak a jestlipak. (English translation counterparts of Czech sentences containing the particles copak and jestlipak.) Časopis pro moderní filologii 98 (1), 228-237.

• • •

Jolanta Šinkūnienė

Vilnius University

jolanta.sinkuniene@flf.vu.lt

Evidential adverbials and stance in English and Lithuanian journalistic discourse: a semantic functional study

Within the last several decades, numerous studies of evidential adverbials in Germanic, Romance, Slavic and Baltic languages confirm semantic versatility and pragmatic multifunctionality of these stance markers in different contexts (Simon-Vandenbergen, Aijmer 2007; Cornillie 2010; Wiemer & Kampf 2012; Usonienė 2015). One of the areas in which writers frequently employ evidential markers to build interpersonal relations and to achieve certain rhetorical goals is journalistic discourse. Evidential markers are very important in journalistic discourse as they point towards the source of information. Sourcing is crucial in the journalistic domain as it "gives us an answer to one of the important questions that readers may ask of a news text: 'to whom can this be attributed?'" (Bednarek 2006: 638).

Research on evidential markers in journalistic discourse focused on their functional variation (Celle 2009a, 2009b; Hennemann 2012), distribution in different subgenres and in newspapers of different political orientation (Hidalgo 2006; Marín 2006; Marín-Arrese 2015). The results suggest interesting trends in the way newspaper discourse is shaped in different cultures as well as confirm the versatile profile of evidential adverbials as stance markers. In most of the cross-linguistic studies the comparative axis is drawn between English vs French or Spanish. The present study aims to compare the functional semantic profile of evidential adverbials in journalistic discourse in English vs Lithuanian in two sub-genres (news reports and opinion articles). A careful contextual analysis is employed in the study to explore how types of evidence (inference, report) and its reliability (high, medium, low) influence the expression of author stance regarding the reported information in the two news subgenres and two languages.

The study is based on a self-compiled comparable corpus of newspaper language; the size of the corpus is roughly 400 000 words. The Lithuanian data comes from the national daily newspaper *Lietuvos Rytas*, whereas the texts in English are from the American daily newspaper *The New York Times*. The study employs quantitative and qualitative methods to account for the frequency and semantic functional distribution patterns of the evidential markers under study.

The preliminary results suggest that there is a tendency for news reports to employ evidential adverbials signaling higher reliability as compared to the opinion articles. The analysis also points towards interesting semantic functional differences between the use of markers in both languages.

References

Bednarek, M. (2006). Epistemological Positioning and Evidentiality in English News Discourse: A Text Driven Approach. Text and Talk, 26 (6), 635–660.

Celle, A. (2009a). Hearsay adverbs and modality. In R. Salkie, P. Busuttil & J. van der Auwera (Eds.), Modality in English theory and description. Berlin, New York: Mouton de Gruyter, 269–293.

Celle, A. (2009b). The intersubjective function of modal adverbs: a contrastive English-French study of adverbs in journalistic discourse. Languages in Contrast, 9 (1), 23-36.

Cornillie, B. (2010). An interactional approach to epistemic and evidential adverbs in Spanish conversation. In G. Diewald & E. Smirnova (Eds.), Linguistic realization of evidentiality in European languages. Berlin, New York: Mouton de Gruyter, 309-330.

Hennemann, A. (2012). The epistemic and evidential use of Spanish modal adverbs and verbs of cognitive attitude. Folia Linguistica, 46 (1), 133-170.

Hidalgo, L. (2006). The expression of writer stance by modal adjectives and adverbs in comparable corpus of English and Spanish newspaper discourse. A. M. Hornero, M. J. Luzón & S. Murillo (Eds.), Corpus linguistics: Applications for the study of English. Bern, Berlin: Peter Lang, 125-140.

Marín, J. I. (2006). Epistemic stance and commitment in the discourse of fact and opinion in English and Spanish: A comparable corpus study. A. M. Hornero, M. J. Luzón & S. Murillo (Eds.), Corpus linguistics: Applications for the study of English. Bern, Berlin: Peter Lang, 141-157.

Marín-Arrese, J. I. (2015). Epistemicity and stance: a cross-linguistic study of epistemic stance strategies in journalistic discourse in English and Spanish. Discourse Studies, 17 (2), 210-225.

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

Usonienė, A. (2015). Non-morphological realizations of evidentiality: The case of parenthetical elements in Lithuanian. In P. Arkadiev, A. Holvoet & B. Wiemer (Eds), Berlin: Mouton de Gruyter, 437-463.

Wiemer, B. & V. Kampf. (2012). On conditions instantiating tip effects of epistemic and evidential meanings in Bulgarian. Slověne: International Journal of Slavic Studies (2), 5-38.

• • •

Workshop 3

Investigating variation in legal discourse

Workshop convenors:
Teresa Fanego (University of Santiago de Compostela) &
Paula Rodríguez-Puente (University of Oviedo)

Donata Berūkštienė
Department of Foreign Language, Literature and Translation Studies, Faculty of Humanities, Vytautas Magnus University, Kaunas, Lithuania
donaber@gmail.com

Structural Types of Lexical Bundles in Court Judgments in English and Lithuanian: a Corpus-Driven Analysis

Legal texts are often criticized for being abstruse and incomprehensible to the general public. The problem is usually with the language used in this kind of texts, i.e. legal language, which is understood as a special-purpose language used by lawmen (e.g. a law student, a lawyer, a judge, a law maker, a law officer, a law researcher, a jurist) in written and spoken texts in the context of law. Legal language is "subject to special syntactic, semantic and pragmatic rules" (Šarcevic 2000: 8). Formulaicity is one of the characteristic features of legal language. Formulaic language, which includes idioms, phrases, collocations, lexical bundles, etc., influences not only the form but also the content of legal texts. However, there has been little research available on the nature of frequently occurring "sequences of three or more words that show a statistical tendency to co-occur" (Biber and Condrad 1999, 183), i.e. lexical bundles, in different genres of legal texts. What is more, existing investigations of lexical bundles in legal texts are based on one language (e.g. Jablonkai 2009, Gozdz-Roszkowski 2011, Breeze 2013) whereas a cross-language comparison of recurrent sequences of word combinations is lacking. To fill the aforementioned gaps, this paper aims at the identification and analysis of structural types of lexical bundles prevailing in court judgments of the Court of Justice of the European Union in English and Lithuanian. As the analysis deals with automatically retrieved multi-word units, the methodological guidelines of corpus linguistics are followed in the course of the investigation of the frequency and structure of lexical bundles. The classification of lexical bundles into structural types is based on the framework suggested by Biber, *et al.* (1999, 2004). For the purpose of this study, two corpora of court judgments have been created. One corpus comprises approximately 1 million words of court judgments in the English language; the other

corpus includes 730 000 words of court judgments in the Lithuanian language. Lexical bundles in this research have been identified by n-gram extraction method using the corpus analysis toolkit AntConc 3.4.4. A concordance program AntPConc has been used to find possible Lithuanian equivalents of the most frequent lexical bundles identified in the English court judgments and to compare the structure of these lexical bundles in both languages.

References

Biber, D., Conrad, S. & V. Cortes. (2004). If You Look at…: Lexical Bundles in University Teaching and Textbooks. Applied Linguistics 25, 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). The Longman Grammar of Spoken and Written English. London: Longman.

Biber, D., & Conrad, S. (1999). Lexical Bundles in Conversation and Academic Prose. In H. Hasselgard & S. Oksefjell (Eds.), Out of Corpora. Amsterdam: Rodopi, 181–190.

Breeze, R. (2013). Lexical Bundles across Four Legal Genres. International Journal of Corpus Linguistics 18 (2), 229-253.

Gozdz-Roszkowski, S. (2012). Patterns of Linguistic Variation in American Legal English: a Corpus-Based Study. Frankfurt am Mein: Peter Lang.

Jablonkai, R. (2009). In the Light of: a Corpus-Based Analysis of Lexical Bundles in Two Eu-Related Registers. WoPaLP 3, 1-27.

Šarčevic, S. (2000). New Approach to Legal Translation. The Hague: Kluwer Law International.

• • •

Doug Biber
Northern Arizona University
douglas.biber@nau.edu

Are law reports an 'agile' or an 'uptight' register? Historical change in the use of phrasal complexity features

Linguistic researchers have often asserted that spoken discourse is real language while written discourse is only an artifact, and that therefore grammatical change emerges exclusively in speech. In contrast, the present authors have argued in a series of previous studies that grammatical change emerges in written (as well as spoken) discourse (e.g., Biber & Gray 2011, 2013, 2016). There are two theoretical assumptions underlying this argument. The first is uncontroversial: that grammatical change emerges in actual language use. The second, though, is unique to the Biber/Gray studies: that written discourse (as well as spoken discourse) is language-in-use, and therefore there

is every reason to believe that the communicative situation of writing should provide the context for grammatical change.

Based on large-scale corpus investigations, the Biber/Gray studies show that these theoretical expectations are borne out. In particular, several types of phrasal complexity features in English have emerged in written discourse over the last 400 years, resulting in changes in both the frequency of use and the grammatical functions of these devices. Academic research writing has been one of the loci for these historical developments, reflecting the peculiar situational characteristics of this register (production circumstances that permit extreme revision and editing of the text, highly informational communicative purposes, and discourse produced by experts for consumption by other specialists). However, there are also systematic differences among academic registers, with science research writing showing the most extreme patterns of linguistic change, while humanities research writing has changed relatively little over the centuries.

The present study investigates whether similar patterns of grammatical change have occurred in the register of law reports. Within the universe of legal registers, law reports are centrally important because they provide official records of judicial decisions which are used to establish legal precedent in later cases (see Fanego et al. 2017). Against the background of our previous research on grammatical complexity, law reports are interesting because they share many communicative characteristics with academic research articles, but they also differ in some potentially important respects: On the one hand, law reports are similar to academic articles in that they incorporate expository explanations and informational descriptions. However, law reports differ in that they also serve an 'operative' function, summarizing the judgement that gives the actual disposition of a legal case. In addition, it might be argued that the communicative characteristics of law reports have changed little over the centuries, in contrast to dramatic changes in the readership and communicative purposes of science research articles. Based on analysis of the CHELAR corpus (see Fanego et al 2017), the present study investigates the competing influence of these factors in relation to the historical development of phrasal complexity features, exploring the extent to which law reports are an 'agile' register (see Hundt & Mair 1999) that is receptive to historical innovations versus an 'uptight' and conservative register that resists historical change.

References

Biber, D., & Gray, B. (2011). Grammar emerging in the noun phrase: The influence of written language use. English Language and Linguistics, 15, 223-250.

Biber, D., & Gray, B. (2013). Being specific about historical change: The influence of sub-register. Journal of English Linguistics, 41, .104-134.

Biber, D., & Gray, B. (2016). Grammatical complexity in academic English: Linguistic change in writing. Cambridge: Cambridge University Press.

Fanego, T., Rodríguez-Puente, P., José López-Couso, M., Méndez-Naya, B., Núñez-Pertejo, P., Blanco-García, C. & Tamaredo, I. (2017). The Corpus of Historical

English Law Reports 1535-1999 (CHELAR): A resource for analysing the development of English legal discourse. ICAME Journal, 41, 53-82.
Hundt, M., & Mair, C. (1999). "Agile" and "uptight" genres: The corpus-based approach to language change in progress. International Journal of Corpus Linguistics, 4, 221-242.

• • •

Giuliana Diani
University of Modena and Reggio Emilia, Italy
giuliana.diani@unimore.it

Lexical bundles in English and Italian legal texts: a cross-linguistic analysis

Research in corpus linguistics have shown that fixed word combinations and formulaic sequences are pervasive in language use (e.g. Altenberg 1993; Moon 1998; Schmitt 2004; Stubbs 2007). Great attention has been devoted particularly to this issue in the field of English for Academic Purposes (e.g. Cortes 2002, 2004; Biber, Conrad & Cortes 2004; Nesi & Basturkmen 2006; Biber 2009; Ädel & Erman 2012; Salazar 2014). However, little attention has been awarded to Language for Specific Purposes studies (e.g. Picht 1990; Bergenholtz & Tarp 1995; Heid 2001). Quantitative analysis of repeated strings of words is an ideal starting point for the exploration of the systematic relation between text and form (Sinclair 2005), but this needs to be related to significant functions, such as for example discourse relations (Siepman 2005). This paper takes into consideration the use of lexical bundles (Biber et al. 1999) in legal texts from a cross-linguistic perspective (English/Italian). Analyses of this linguistic area of study are still relatively rare in legal language (Jablonkai 2010), and often limited to English (Gozdz-Roszkowski 2011; Breeze 2013; Kopaczyk 2013). The purpose of this study is to identify and analyse the most typical lexical bundles occuring in English and Italian legal texts, so as to discover some common ground in the use of these linguistic features. Looking at them will contribute to shed some light on similarities and/or differences in their form and function across the two languages under investigation.

References

Ädel, A. & Erman B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: a lexical bundles approach. English for Specific Purposes, 31 (2), 81-92.

Altenberg, B. (1993). Recurrent word combinations in spoken English. In J. D'Arcy (Ed.), Proceedings of the Fifth Nordic Association for English Studies Conference. Reykjavik: University of Iceland, 17-27.
Bergenholtz, H. & Tarp, S. (1995). Manual of Specialised Lexicography. Amsterdam: John Benjamins.
Biber, D. (2009). A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing. International Journal of Corpus Linguistics, 14 (3), 275-311.
Biber, D., Conrad, S. & Cortes, V. (2004). If you look at …: lexical bundles in university teaching and textbooks. Applied Linguistics, 25 (3), 371-405.
Breeze, R. (2013). Lexical bundles across four legal genres. International Journal of Corpus Linguistics, 18 (2), 229-253.
Cortes, V. (2002). Lexical bundles in freshman composition. In R. Reppen, S. Fitzmaurice & D. Biber (Eds.), Using Corpora to Explore Linguistic Variation. Amsterdam: John Benjamins, 131-146.
Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: examples from history and biology. English for Specific Purposes, 23 (4), 397-423.
Goźdź-Roszkowski, S. (2011). Patterns of Linguistic Variation in American Legal English. A Corpus-Based Study. Peter Lang: Frankfurt am Main.
Heid, U. (2001). Collocations in sublanguage texts: extraction from corpora. In S. E. Wright & G. Budin (Eds.), Handbook of Terminology Management. Amsterdam: John Benjamins, 788-808.
Jablonkai, R. (2010). English in the context of European integration: a corpus-driven analysis of lexical bundles in English UE documents. English for Specific Purposes, 29 (4), 253-267.
Kopaczyk, J. (2013). The Legal Language of Scottish Burghs. Standardization and Lexical Bundles (1380-1560). Oxford: Oxford University Press.
Moon, R. (1998). Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford: Clarendon Press.
Nesi, H. & Basturkmen, H. (2006). Lexical bundles and discourse signaling in academic lectures. International Journal of Corpus Linguistics, 11 (3), 283-304.
Picht, H. (1990). LSP phraseology from the terminological point of view. Terminology Science and Research, 1 (1-2), 33-48.
Salazar, D. (2014). Lexical Bundles in Native and Non-Native Scientific Writing: Applying a Corpus-Based Study to Language Teaching. Amsterdam: John Benjamins.
Schmitt, N. (2004). Formulaic Sequences: Acquisition, Processing, and Use. Amsterdam: John Benjamins.
Siepmann, D. (2005). Discourse Markers across Languages. A Contrastive Study of Second-Level Discourse Markers in Native and Non-Native Text with Implications for General Pedagogy and Lexicography. London / New York: Routledge.
Sinclair, J. McH. (2005). Language as a string of beads: discourse and the m-word: In E. Tognini-Bonelli & G. Del Lungo Camiciotti (Eds.), Strategies in Academic Discourse. Amsterdam: John Benjamins, 163-168.

Stubbs, M. (2007). An example of frequent English phraseology: distribution, structures and functions. In R. Facchinetti (Ed.), Corpus Linguistics 25 Years On. Amsterdam: Rodopi, 89-105.

• • •

Teresa Fanego
University of Santiago de Compostela
teresa.fanego@usc.es

Paula Rodríguez-Puente
University of Oviedo
rodriguezppaula@uniovi.es

Investigating variation in legal discourse: An overview

Investigation into law and language is extensive. For linguists, interest in the field emanates from the recognition of law language as a fruitful source of data for linguistic analysis and for testing theories about language and discourse. For researchers in other disciplines, such as psychology, sociology and anthropology, language as it operates within the legal system serves as a vehicle for understanding the legal process itself but also "as a means for investigating psychological processes, societal interactions, or cultural traits" (Schane 2006: 4). The approach adopted in this workshop is primarily linguistic. We are therefore concerned with issues such as the following:

(1) Existing typologies of legal discourse, which are based mostly on contemporary usage, account for the heterogeneity of legal language by distinguishing legal texts in terms of:

- their degree of formality – frozen, formal, consultative and casual are labels often used in this connection; see Danet (1980);
- their communicative purpose – whether this is academic, juridical or legislative; see Bhatia (1987);
- their mode (written vs oral).

Questions that emerge here pertain, first, to whether such categorizations can also be fruitfully applied to the analysis of legal discourse in earlier stages of English. Secondly, to the availability, or non availability, of databases adequate to carry out such an analysis.

To address both questions, the workshop surveys recent developments in the compilation of electronic corpora containing legal documents of various kinds, both synchronic and diachronic. Among these, the following deserve special mention: American Law Corpus (Goźdź-Roszkowski 2011); Old Bailey Corpus (Huber et al. 2012); Corpus of Early Modern English Statutes 1491-1707 (Lehto 2013); CHELAR -

Corpus of Historical English Law Reports 1535-1999 (Rodríguez-Puente, Fanego et al. 2016; Fanego, Rodríguez-Puente et al. 2017).

(2) Both classic (e.g. Mellinkoff 1963, Crystal & Davy 1967, Gustaffsson 1975, Finegan 1982, Bhatia 1993) and recent (Scotto di Carlo 2015) treatments of the language and law interface have drawn attention to various lexical, morphosyntactic and discoursal features that are claimed to be inextricably linked to the language of the law: use of Norman words that have not found their way into general currency, heavy use of compound adverbs such as hereof, whereof, hereinafter, binomial and multinomial expressions (e.g. within Singapore or elsewhere), lexical bundles and phraseological units (e.g. the benefit of, it is clear that, on the basis that), intricate patterns of coordination and subordination, impersonal style and frequent use of passive constructions, conditional constructions, etc. Yet exemplification of all such features tends to draw heavily on legislative texts such as acts of parliament and statutory instruments, these being, in fact, the only legal writings usually discussed in the relevant literature. The workshop, therefore, also addresses the question of internal variation across legal genres: how and to what extent do legal genres differ from, or are similar to, each other?

(3) Other important dimensions of variation in legal discourse pertain to diachronic variation (how does the current legal language, or languages, differ from the historic one?) and to so-called external variation (how does legal language differ from other registers, or from other languages for special purposes?). The development of Multi-Dimensional analysis from the 1990s onwards (Biber 1988, 1995, 2001, 2013, etc.), and the recent advances in the compilation of synchronic and diachronic corpora of legal English mentioned under (1) above, have now provided the resources enabling researchers to carry out corpus-based comprehensive analyses of variation in legal discourse over time, as well as relative to other genres and registers.

References

Bhatia, V. K. (1987). Language of the law. Language Teaching, 1987, 227-234.

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. (1995). Dimensions of Register Variation: A Cross-Linguistic Perspective. Cambridge: Cambridge University Press.

Biber, D. (2001). Dimensions of variation among 18th century registers. In H.-J. Diller & M. Görlach (Eds.), Towards a History of English as a History of Genres. Heidelberg: C. Winter, 89-110.

Biber, D. (2013). Multi-Dimensional analysis. A personal history. In T. Berber Sardinha & M. Veirano Pinto (Eds.), Multi-Dimensional Analysis, 25 Years On. A Tribute to Douglas Biber. Amsterdam: John Benjamins, xxix-xxxviii.

Crystal, D. & D. Davy. (1969). Investigating English Style. London: Longman.

Danet, B. (1980). Language in the courtroom. In H. Giles, P. Smith & W. P. Robinson (Eds.), Language: Social and Psychological Perspectives. Oxford: Pergamon, 367-376.

Fanego, T. & Rodríguez-Puente, P. et al. (2017). The Corpus of Historical English Law Reports 1535-1999 (CHELAR): A resource for analysing the development of English legal discourse, ICAME Journal 41, 67-96.
Finegan, E. (1982). Form and function in testament language. In R. J. Di Pietro (Ed.), Linguistics and the Professions. Proceedings of the Second Annual Delaware Symposium on Language Studies. Norwood, N. J.: Ablex, 113-120.
Goźdź-Roszkowski, S. (2011). Patterns of Linguistic Variation in American Legal English. A Corpus-Based Study. Frankfurt am Main: Peter Lang.
Gustaffsson, M. (1975). Some Syntactic Properties of English Law Language. Turku: Department of English, University of Turku.
Huber, M. & Nissel, M. & Maiwald, P. & Widlitzki, B. (2012). The Old Bailey Corpus. Spoken English in the 18th and 19th Centuries. http://www.uni-giessen.de/oldbaileycorpus; http://fedora.clarin-d.uni-saarland.de/oldbailey/
Lehto, A. (2013). Complexity and genre conventions. Text structure and coordination in Early Modern English proclamations. In A. H. Jucker, D. Landert, A. Seiler & N. Studer-Joho (Eds.), Meaning in the History of English. Words and Texts in Context. Amsterdam: John Benjamins, 233-256.
Mellinkoff, D. (1963). The Language of the Law. Boston, MA: Little, Brown and Company.
Rodríguez-Puente, P. & Fanego, T. & López-Couso, M. J. & Méndez-Naya, B. & Núñez-Pertejo, P. (2016). Corpus of Historical English Law Reports 1535-1999 (CHELAR). University of Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization, Department of English and German.
Schane, S. (2006). Language and the Law. London: Continuum.
Scotto di Carlo, G. (2015). Diachronic and Synchronic Aspects of Legal English: Past, Present, and Possible Future of Legal English. Newcastle upon Tyne: Cambridge Scholars Publishing.

• • •

Nicholas Groom
University of Birmingham
n.w.groom@bham.ac.uk

British patents of invention 1711–2011: A corpus-based diachronic genre analysis

This paper has three interrelated aims. The first aim is to present a novel corpus-based methodology for the diachronic analysis of generic structure. This methodology follows Biber and Conrad (2009) in regarding generic structure as being principally marked by overt textual features such as section headings, fixed phrases, special for-

matting and so on. Essentially, the analysis involves reducing the generic structure configurations of individual exemplar texts to simple code strings that can be processed using adaptations of standard corpus analysis techniques.

The second aim of the paper is to present the results of an empirical study which used this methodology to identify and study changes in the generic structure of a corpus of British patent specification texts between 1711 and 2011. The corpus was compiled from the complete collection of over 2 million historical UK patent documents held at The British Library, and (for more recent texts) from the European Patent Office's *Espacenet* online patent search interface (http://www.epo.org/searching-for-patents/technical/espacenet.html#tab1). On the basis of this analysis I will argue that there have been five major transformations in the structural form of the British patent specification genre in its three hundred years of continuous existence. I will interpret these generic changes in social and functional terms, showing how they can be related to concurrent changes in intellectual property law and its conceptual underpinnings, to developments in science and technology, to the growth of manufacturing industry and other forms of commercial activity during the period, and to broader developments in British society and politics as a whole.

The third aim of the study is to discuss the implications of my empirical analysis of patents for a current theoretical question in diachronic genre studies: is the process of genre change best understood in Darwinian terms, as a matter of constant and gradual evolution (e.g. Gross et al 2002), or is it better understood in Kuhnian terms, as a series of relatively stable periods of activity punctuated by occasional and abrupt revolutionary shifts (e.g. Berkenkotter 2009)? Although the results of my study lean more towards a revolutionary than an evolutionary account of genre change, I will caution against an overly literal Kuhnian interpretation of my data. I will also suggest that the aptness of an evolutionary or a revolutionary interpretation of the results of a diachronic genre analysis may also depend on the function and status of the genre in the society in which it operates, and on the level of the genre analysis itself.

References

Berkenkotter, C. (2009). Patient tales: Case histories and the uses of narrative in psychiatry. Columbia, SC: University of South Carolina Press.

Biber, D. & Conrad, S. (2009). Register, genre, and style. Cambridge: Cambridge University Press.

Gross, A. G., Harmon, J. E., & Reidy, M. (2002). Communicating science. The scientific article from the 17th century to the present. Oxford: Oxford University Press.

• • •

Anu Lehto
University of Helsinki
anu.lehto@helsinki.fi

The representation of authorities and citizens in Acts of Parliament in 1800–2000

Legal texts have the authority to pose rights and obligations to less powerful recipients; the laws reflect existing values in society but at the same time they form new belief systems, shaping the portrayals of different groups of people. Social representations are shared assumptions about people, and they affect how the speakers perceive particular phenomena (Burr 2002: 106). The paper studies the representations of authorities and citizens in British Acts of Parliament from the nineteenth and twentieth centuries. There have been few previous studies on the portrayals of social groups in historical material (e.g. Bäcklund 2006), and earlier research has not considered the representation of the officials or lay people in historical national legislation.

The data is a self-compiled diachronic corpus of late modern acts retrieved from the National Archives of the UK government, and the study takes advantage of collocation analysis. Collocations reveal lexical elements that appear together more often than random word combinations, and the collocates affect the meaning of the node (e.g. McEnery & Hardie 2012). The analysis includes collocates that extend up to five words from the node, and the study identifies groups of collocates with the same semantic meaning (semantic preference) and pays attention to evaluative meanings associated with the social groups (semantic prosody) (Sinclair 2004: 31).

The study is based on genre analysis and historical pragmatics. Genres can form different representations of groups of people, depending on the audience and purpose of texts. Consequently, the attitudinal meaning of collocates can vary according to genres (see Hunston 2007). Collocations and other formulaic wordings emerge when the same communicative situations and goals are repeated in a community (Croft 2000), and these routinised combinations can spread to new contexts (Hoey 2005). Historical pragmatics further considers texts in their sociohistorical contexts, noting possible diachronic changes (Taavitsainen & Jucker 2010). In the Victorian era, legislation aimed especially to improve living conditions of the citizens and to treat people more equally, which can be reflected in the collocates (see Cornish et al. 2010, Rees 2001).

The study shows that the acts form notably dissimilar representations of the authorities and citizens. The crown is addressed in a respective manner and the collocates underline the possessions of the monarch. The most common word that refers to citizens is *person* but the texts define more specific groups as well: the acts single out persons, for instance, by the collocates *idle* and *disorderly*, which refer to vagabonds and people who commit crimes. The citizens are further often collectively referred to as *the Majesty's subjects*, which emphasises their less powerful status. The authorities include various officials, and they are regularly defined through administrative areas such as *parishes*.

References

Bäcklund, I. (2006). Modifiers describing men and women in nineteenth-century English. In M. Kytö, M. Ryden & E. Smitterberg (Eds.), Nineteenth-century English: Stability and Change. Cambridge: Cambridge University Press, 17–55.

Burr, V. (2002). The person in social psychology. New York: Psychology Press.

Cornish, W., Anderson, J. S., Cocks, R., Lobban, M., Polden, P. & Smith, K. (2010). The Oxford History of the Laws of England, 1820–1914. Vols. 11 and 13. Oxford: Oxford University Press.

Croft, W. (2000). Explaining Language Change: An Evolutionary Approach. London: Longman.

Hoey, M. (2005). Lexical Priming: A New Theory of Words and Language. New York: Routledge.

Hunston, S. (2007). Semantic Prosody Revisited. International Journal of Corpus Linguistics, 12 (2), 249–268.

McEnery, T. & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press.

Rees, R. (2001). Poverty and Public Health: 1815–1948. Oxford: Heinemann.

Sinclair, J. (2004). Trust the Text: Language, Corpus and Discourse. London: Routledge.

Taavitsainen, I. & Jucker, A. H. (2010). Trends and Developments in Historical Pragmatics. In A. H. Jucker, & I. Taavitsainen (Eds.), Historical Pragmatics. Berlin and New York: Walter de Gruyter, 3–30.

• • •

Randi Reppen
Northern Arizona University
randi.reppen@nau.edu

Meishan Chen
Northern Arizona University
mc2544@nau.edu

A comparison of lexical bundles in spoken courtroom language across three periods

This paper explores diachronic and register variation in spoken courtroom language using lexical bundles (Biber et. al. 1999). Building on the work of Culpeper and Kytö (2010; pp. 103 - 141) that described lexical bundles in Present Day English (PDE) trials (126,271 words) and Early Modern English (EME) trials (252,607 words), we use the same methods to explore the 1994 trial of O. J. Simpson (220,552 words). Like Culpeper and Kytö, we identified the fifty most frequent bundles in the O. J. Simpson trial and

found twelve bundles that were shared across the three corpora. Four bundles only occurred in PDE (*in this case*; *the fact that*; *at the time*; *that there was*), and three only occurred in EME (*did you see*; *out of the; I did not*). Five bundles were shared across the three corpora (*at that time; what did you; one of the; there was a; I don't know*). Eight of these shared bundles occurred at relatively the same rate of occurrence in the three trials (PDE, EME and OJ).

After comparing the lexical bundles and their functions to Culpeper and Kytö, we divided the OJ Simpson trial into three sub-registers that reflect different contexts: Opening Statement - when lawyers present the situation from their perspectives; Direct Examination - when witnesses are examined 'cooperatively' by lawyers; and Cross Examination - when witnesses are examined by an adversarial lawyer. The bundles in these three sub-registers directly reflect these different situations. The most frequent bundles in the Opening Statement serve to contextualize and introduce the trial. Frequent bundles such as, *evidence will show, you will hear*, and *you will see* are sensory related and are addressed to the jury to prepare them for what they will experience. Bundles in the Direct Examination typically reflect questions designed to elicit information from witnesses that supports the lawyer's position (e.g., *what did you*; *did you do*; *can you tell us*). In contrast, Cross Examination bundles are not as 'soft' as those in the Direct Examination. They are more challenging and therefore elicit expressions of uncertainty (e.g., *is that correct*; *I'm not sure*; *I don't know*).

The results of these analyses show the influence of situational similarities across time, not only in the bundles used, but also the functions of the bundles. The results also show that exploring sub-registers can provide a picture of how language varies due to situational and functional goals.

References

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). The Longman Grammar of Spoken and Written English. London: Longman.

Culpeper, J. & Kytö, M. (2010). Early Modern English Dialogues: Spoken Interaction as Writing. Cambridge: Cambridge University Press.

• • •

Paula Rodríguez-Puente
University of Oviedo
rodriguezppaula@uniovi.es

Are law reports a hybrid text type? Exploring the oral/literate continuum in the Corpus of Historical English Law Reports, 1535-1999

This presentation examines the language of law reports from a synchronic and diachronic perspective, based on the recently released *Corpus of Historical English Law Reports, 1535-1999* (Rodríguez-Puente, Fanego et al. 2016).

In English common law, reports are records of judicial decisions which are "cited by lawyers and judges for their use as precedent in subsequent cases" (*Encyclopædia Britannica Online* s.v. *law report*). Historically, judicial decisions and custom are the most important ways in which the common law has been built up, and hence they "still play a significant role as they interpret parliamentary law and fill in the gaps where there is no statute law" (Kearns 2007: 9).

The aim of the paper is twofold. On the one hand, I intend to focus on, and identify the properties of law reports that had led some scholars to define them as 'hybrid' texts, fulfilling both prescriptive (normative) and descriptive (non-normative) functions and containing both expository (more objective) and operative (more subjective) linguistic features (see, among others, Šarčević 2000: 11, Tiersma 1999: 139-141, Williams 2007: 28-29). On the other hand, the paper looks at how judicial decisions as a text type have evolved linguistically from the early sixteenth century to the present day, especially after the Incorporated Council of Law Reporting for England and Wales (ICLR) became established as the only authorised publisher of the official series of law reports in 1865 (see Fanego, Rodríguez-Puente, et al. 2017).

For these purposes, the analysis takes as its point of departure the set of language features that, in terms of Biber's well-known multi-dimensional typology of texts (1988), can be argued to be most distinctive of the oral/literate continuum. In connection with this, a preliminary study conducted so far indeed suggests that law reports differ from other kinds of legal documents, such as Acts of Parliament and declarations, in that they are partially narrated in the first person and make extensive use of vocative expressions (e.g. *my lords*). Yet, at the same time, they often exhibit features typical of formal, written texts, such as complex sentence structure (e.g. […] *yielding and paying as in the indenture; which indenture of lease is found in haec verba*), academic, learned vocabulary, such as Romance nominalisations (e.g. *maintenance*, *corroboration*), and formal linking devices (e.g. *forthwith*, *hereinafter*, *hereinbefore*, *herewith*, *thereafter*, *thereof*, *thereon*, *thereunder*, *whereof*), among others. From a diachronic perspective, the analysis also reveals that the earlier reports differ greatly from the most recent ones both in content and structure. Thus, the lower frequencies of first and second person pronouns in the earlier reports suggest that these are marked by a more detached and objective style, whereas later texts show a greater degree of involvement.

These and other related issues will be addressed in this paper with the aim of shedding new light on the linguistic features of law reports both synchronically and diachronically, as well as in comparison with other legal documents.

References

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Corpus of Historical English Law Reports 1535–1999 (CHELAR). 2016. Compiled by Rodríguez-Puente, P., Fanego, T., López-Couso, M.J., Méndez-Naya, B. & Núñez-Pertejo, P. University of Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization, Department of English and German.

Encyclopædia Britannica Online. (http://www.britannica.com)

Kearns, M. (2007). Legal English. Madrid: Colex.

Fanego, T., Rodríguez-Puente, P., López-Couso, M.J., Méndez-Naya, B., Núñez-Pertejo, P., Blanco-García, C. & Tamaredo, I. 2017. The Corpus of Historical English Law Reports 1535-1999 (CHELAR): A resource for analysing the development of English legal discourse. ICAME Journal, 41, 67-96.

Šarčević, S. (2000). New Approach to Legal Translation. The Hague: Kluwer Law International.

Tiersma, P. M. (1999). Legal Language. Chicago, IL: The University of Chicago Press.

Williams, C. (2007). Tradition and Change in Legal English. Verbal Constructions in Prescriptive Texts (2nd ed.). Bern: Peter Lang.

• • •

Workshop 4

Back to text – Contextualizing corpus data in historical and variationist English linguistics

Workshop convenors:
Kristin Bech (University of Oslo) &
Ruth Möhlig-Falke (Heidelberg University)

Lynn Anthonissen
University of Antwerp
lynn.anthonissen@uantwerpen.be

Peter Petré
University of Antwerp
peter.petre@uantwerpen.be

The EMMA corpus: balancing big data and contextualization

With ever increasing corpus sizes, much of the valuable contextual information that has been key to small-scale historical linguistic and variationist studies of English has been lost. While a qualitative contextual analysis is indispensable for the sociocultural dynamics of language use, advanced quantitative analyses do require large databases, in particular when dealing with low frequency phenomena. This area of tension raises two pertinent issues for corpus linguists:

1. Which criteria must be met in order to ensure representativeness of large corpora?
2. To what extent can *big data* corpora still take into account the sociohistorical context in which the texts were produced?

In this talk, we demonstrate how these issues are tackled within the context of a collaborative research project on Early Modern English. Situated at the intersection of historical sociolinguistics and corpus linguistics, the Mind-bending Grammars project aimed to establish a new longitudinal corpus (called *Early Modern Multiloquent Authors* [*EMMA*]) that sets high standards for both qualitative and quantitative corpus requirements.

Existing corpora of Early Modern English are varied in scope and size. Specialized corpora are rigorously compiled and representative of specific language uses, but

generally small. Examples include the corpora of Early Modern Correspondence[1], English Dialogues[2], and Early English Medical Writing[3]. Extensive digitalization projects such as Early English Books Online (EEBO) provide digitized editions of writings by all British authors between 1600-1800. Yet they are unstructured databases rather than real corpora. Well-established multi-purpose corpora, such as the PENN-parsed PPCEME2[4], the Helsinki Corpus[5] and ARCHER[6], approach a high degree of balance but are relatively small in size. The EMMA corpus fills a gap by being a large-scale specialized corpus, with a size of about a 100 million words covering a social network of 50 17th-century London-based writers. The body of texts was mainly collected from the EEBO and ECCO databases following an extensive author selection process. The set of criteria to be fulfilled by the prospective authors (including, among others, a long career, a demonstrable link with London and social, political and stylistic network connections) ensures that EMMA is a representative corpus of Early Modern English as written and/or spoken by the intellectual elite.

Of course, the sheer size of a corpus such as EMMA prevents a profound philological analysis of the sociocultural context in which the individual texts were written. However, by efficient use of other available resources (e.g. retrieving information from large databases such as the Oxford Dictionary of National Biography[7]) we were able to establish a rich metadata database, which is integrated in the corpus query and annotation tool (*CosyCat*[8]). Other measures included rigorous date and authorship verification and genre classification. In all these aspects, the EMMA corpus provides a qualitative edge over the large databases that are currently being used in Digital Humanities and 'big data' corpus studies, while at the same time it significantly broadens the scope of historical sociolinguistic studies.

References

[1] http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/
[2] http://www.helsinki.fi/varieng/CoRD/corpora/CED/
[3] http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/
[4] https://www.ling.upenn.edu/hist-corpora/
[5] http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/
[6] http://www.projects.alc.manchester.ac.uk/archer/
[7] http://www.oxforddnb.com/
[8] https://github.com/emanjavacas/cosycat/blob/master/README.md

• • •

Antonette Healey
University of Toronto
healey@doe.utoronto.ca

The Enabling Archive: Old English Textual Transmission and the Dictionary of Old English Corpus (DOEC)

The body of surviving Old English encompasses a rich diversity of records, written on parchment, carved in stone or bone, or inscribed in jewelry. These texts fall into several categories: poetry (6%), prose (69%), glosses and glossaries to Latin texts (25%), and inscriptions (<1%) in both the runic and Latin alphabets. In the prose in particular, there is a wide range of texts: saints' lives, sermons, biblical translations, penitential writings, laws, charters, wills, records of various sorts (of manumissions, land grants, land sales, land surveys), chronicles, a set of tables for computing the moveable feasts of the Church calendar and for astrological calculations, medical text, prognostics (the Anglo-Saxon equivalent of the horoscope), charms (such as those for a toothache or an easy labor), and even cryptograms. The Dictionary of Old English Project at the University of Toronto has digitized this corpus of material into a comprehensive database. The DOEC comprises at least one copy of each surviving Old English text, and sometimes more than one copy if of interest because of dialect, date, point of view, etc. Today the DOEC consists of 3060 texts, ca. three million words of Old English to which approximately another million words of Latin is attached. It occupies 49 MB, or in literary terms its size is about five times the size of the Collected Works of Shakespeare.

As the DOEC is a closed corpus of manageable size, it can be an invaluable aid for linguistic research, for it is a rich source of early information on the story of English. However, the information derived from it must constantly be assessed in relation to its natural environment, that is, the full context out of which it arises. This workshop intends to demonstate how the DOEC is an enabling tool, exploring the earliest forms of the English language through an investigation of the text types included in the database, and the corresponding freedom in / constraints on interpretation that follow.

References

Healey di Paolo, A., Wilkin, J. P. & Xiang, X. (2009). Dictionary of Old English Web Corpus. Toronto: Dictionary of Old English Project.

Healey di Paolo, A., Holland, J., McDougall, D., & McDougall, I. (2009). Dictionary of Old English Corpus on CD-ROM. TEI-P5 conformant version. Toronto: Dictionary of Old English Project.

• • •

Magnus Huber
University of Giessen, Germany
magnus.huber@anglistik.uni-giessen.de

The Database of Early Pidgin and Creole Texts (DEPiCT): assessing the reliability of early descriptions of contact languages

The Database of Early Pidgin and Creole Texts (DEPiCT) assembles early attestations and descriptions of contact languages. The texts are annotated and made searchable online. The considerable advantages of collecting such texts in one place extend to the following:

- DEPiCT offers a more complete overview and more comprehensive documentation of the development and history of individual contact languages.
- By allowing direct comparison between texts, DEPiCT makes it possible to evaluate the reliability of early sources.

The annotation of linguistic data includes the morphologically segmented text, the orthographic lexifier equivalence, the source language, the gloss, POS tagging and a free translation. More importantly in the present context, there is also contextual and sociolinguistic annotation (e.g. for information on the location of the utterance, descriptions of the language ecology of the different localities, the domains of language use and language attitudes, as well as socio-biographical speaker information). The DEPiCT annotation facilitates systematic research on the emergence and use of contact languages. It also helps to correlate object language with sociolinguistic parameters, thus helping to contextualize language data.

This talk will focus on the challenges in compiling the database. Because of the nature and uneven availability and quality of early descriptions of contact languages, DEPiCT is not a balanced corpus in the classic sense. Rather, we collect what we can and use standardized annotation categories to make the texts maximally searchable and comparable. After a glance at the kinds of texts included in DEPiCT (e.g. travel accounts, narratives by missionaries and traders, official documents by colonial administrators), I will address the question of the reliability of these texts, highlighting questions such as prejudices and the political and/ or cultural agendas of the authors, their linguistic and observational abilities, the time lapse between observation and composition/ publication of the text, the reliability of different genres, the question of plagiarism, as well as the itinerary of the author and time spent in individual localities.

References

The Database of Early Pidgin and Creole Texts (DEPiCT). http://www.uni-giessen.de/faculties/f05/engl/ling/staff/professors/mhuber/depict

• • •

Magnus Huber
University of Giessen, Germany
magnus.huber@anglistik.uni-giessen.de

The communicative and textual context of the Old Bailey Corpus 1720-1913

This presentation focusses on the communicative and textual context of the Old Bailey Corpus (*OBC 2.0*), a corpus of speech-related texts covering the 1720-1913 period and based on the proceedings of London's central criminal court (http://www.oldbaileyonline.org). The 24 million words corpus of speech-related Late Modern English texts is hosted at the CLARIN-D Fedora Commons Repository of Saarland University and is freely accessible for non-commercial research (https://fedora.clarin-d.uni-saarland.de/cqpweb/obc).

A particular strength of the *OBC* is that it includes a very high number of speakers and therefore constitutes a fairly representative sample of spoken, rather formal Late Modern English in the courtroom setting. Every speaker turn has detailed annotation for sociobiographical (gender, social class, age), pragmatic (role in the trial) and textual variables (the shorthand scribe, printer and publisher of individual Proceedings), thus including crucial parameters of the communicative situation. The combination of sheer size and rich annotation make the *OBC* the largest diachronic collection of spoken English with this detail of utterance-level sociolinguistic annotation. The corpus is a valuable resource for multivariate analyses in historical linguistics in general, and historical sociolinguistics and pragmatics in particular, enabling researchers to correlate linguistic change and structural variability with the social context.

The fact that the proceedings of the Old Bailey were taken down in shorthand makes them a reasonably close representation of what was said in the courtroom. Nevertheless, scribes, printers, publishers and the constraints of the printed medium did of course act as linguistic filters between what was actually said and what made its way onto the printed page. In an attempts to contextualize the texts in the *OBC* and assess their reliability, I will discuss the nature of these filters and their effects on the representation of the speech events in the courtroom. The discussion will be based on linguistic evidence within the corpus, a comparison of individual trials in the *OBC* with alternative accounts, testimonies of the scribes about their scribal practices, and the shorthand system they used. The communicative setting of the courtroom and the roles of individual participants in the proceedings also had an important influence on what individuals could and could not say. Another important factor to be taken into consideration is the fact that the genre of the trial proceeding changed during the 18th and 19th centuries, from a privately sold commercial venture favouring rather informal language to an official document of what was said in court and more formal language.

References

The Old Bailey Corpus (OBC). https://fedora.clarin-d.uni-saarland.de/cqpweb/obc

• • •

Ursula Lenker
University of Munich
Ursula.Lenker@lmu.de

Reclaiming Space - The Manuscript Context of Old English Biblical Texts in the Helsinki Corpus

This paper, which the workshop convenors asked to address the theme „from manuscripts to corpus", will focus on some of the biblical and homiletic material selected for the Old English parts of the *Helsinki Corpus of English Texts* (1991). In an attempt at „reclaiming space", I will show how research on Old English information structure can profit from inspecting Old English texts in the design context of their respective manuscripts, in their layout in pages, lines or columns.

Test cases will consider visual means of highlighting episode structure, by, for instance, the use of colours, rubrics and larger initials in different manuscripts of the *West-Saxon Gospels* (COWSGOSP), and, in particular, the value of different textual arrangements of the Latin and Old English texts in the interlinear glosses to the *Lindisfarne* and *Rushworth Gospels* (COLINDIS, CORUSHW).

Specific emphasis will be placed on the functions of Old English *þa* in Farman's glosses to the gospel according to Matthew in the *Rushworth Gospels*, in particular in verses without a Latin lemma such as *tunc*, *cum* or *autem* and it will be tested whether these insertions of *þa* are a signal of idiomatic discourse structuring in Old English.

• • •

Markus Manfred
English Department, The University of Innsbruck
manfred.markus@uibk.ac.at

The Heritage of Late Modern English Dialects in Middle English (based on EDD Online)

While it is a truism that language, and dialectal English in particular, are persistent historical phenomena, this has hardly ever been proved except by eclectic examples. This paper provides cumulative evidence of the survival of Middle English in the dialects of the 18th and 19th centuries. The evidence can be given thanks to the availability since 2016 of *EDD Online*, the digitised version of Wright's comprehensive *English Dialect Dictionary* (1898-1905).

After a short introduction to the interface of this new platform, created at the University of Innsbruck, with Manfred Markus as project director, this paper pursues two aims: (1) it demonstrates the usability of *EDD Online* as a tool for tracing lexical survival from OE and ME to (Late) Modern English, mainly due to the applicability of a time filter; (2) it discusses the various search routines provided by the Innsbruck interface in view of the heritage issue. The main questions raised are: *what* survived, and *where* can the survivals be traced? As to the substance of survival (*what*), this paper's survey ranges from literary specification, using the Gawain works as a test case, to types of word formation, and from spelling and pronunciation to syntax. The question of the *where* of survival familiarises us with Wright's most complex areal attribution of dialectal features, from hundreds of *counties* or their subdivisions to the larger *regions* of the UK and to English-speaking *nations* worldwide.

The paper pairs a linguistic and variationist interest in historical English with a methodological corpus-linguistic interest in Wright's *English Dialect Dictionary* and its usability as a corpus. In the face of the linguistic expertise brought to bear on dialectal raw data, both by Wright with his *EDD* and the Innsbruck team with their interface parameters, our concept of a "corpus", the paper suggests, has to be redefined in the direction of an "ordered corpus" – a collection of data logistically well sorted like the articles of a modern supermarket. The position of the corpus analyst, the consumer, as it were, is, thus, greatly improved.

References

http://eddonline-proj.uibk.ac.at (= Innsbruck platform of EDD Online)

Markus, M. (2010a). Diminutives in English standard and dialects: A survey based on Wright's English Dialect Dictionary. In M. Markus, C. Upton & R. Heuberger (Eds.), Joseph Wright's English Dialect Dictionary and beyond: Studies in Late Modern English dialectology. Frankfurt am Main etc.: Peter Lang, 111-129.

Markus, M. (2010b). As drunk as muck. The role and logic of similes in English dialects on the basis of Joseph Wright's English Dialect Dictionary. Studia Neophilologica, 82, 203-216.

Markus, M. (2012). The complexity and diversity of the words in Wright's English Dialect Dictionary. In M. Markus, Y. Iyeiri, R. Heuberger & E. Chamson (Eds.), Middle and Modern English corpus linguistics: A multi-dimensional approach. Amsterdam and Philadelphia: John Benjamins Publishing Company, 209-224.

Markus, M . (2014). The pattern to be a-hunting from Middle to Late Modern English. Towards extrapolating from Wright's English Dialect Dictionary. In K. Davidse, C. Gentens, L. Ghesquière & L. Vandelanotte (Eds.), Corpus Interrogation and Grammatical Patterns. Amsterdam: John Benjamins Publishing Company, 57–80.

Wright, J. (1898-1905). The English Dialect Dictionary. 6 vols. Oxford: Henry Frowde.

• • •

Terttu Nevalainen
University of Helsinki
terttu.nevalainen@helsinki.fi

Contextualizing the Corpus of Early English Correspondence (c. 1400–1800)

A good deal of variationist sociolinguistic research has been carried out using the Corpus of Early English Correspondence (CEEC). The original version of the corpus covers the time span from 1410 to 1680 and its Extension (CEECE) brings it to 1800, the whole corpus totalling 5.2 million words.

The corpus comes in various formats and with metadata of different kinds. The first published version was a half-a-million-word sampler of the plain text original (CEECS, 1999). The original version has also been published with grammatical annotation (PCEEC, 2006). This version consists of 2.2 million words.

The PCEEC attaches to each sentence sociolinguistic metadata about the letter (date, authenticity, writer–recipient relationship) and the writer and the recipient (name, gender, date of birth, age). There is also a separate 'associated information file' (AIF) that includes additional information on the letters and correspondents. The parsed version for the extract "nor the commyssion for the pease I never harde of", for example, comes out as:

```
((METADATA (AUTHOR NICHOLAS_BACON_II:MALE:BROTHER:1543:26)
        RECIPIENT NATHANIEL_BACON_I:MALE:BROTHER:1546?:23?)
        LETTER BACON_001:E1:1569:AUTOGRAPH:FAMILY_NUCLEAR))
  (IP-MAT (CONJ nor)
       (NP-1 (D the) (N commyssion)
            (PP (P for)
                (NP (D the) (N pease))))
       (NP-SBJ (PRO I))
       (ADVP-TMP (ADV never))
       (VBD harde)
       (PP (P of)
          (NP *ICH*-1))
       (. .)) (ID BACON,I,7.001.5))
```

More comprehensive metadata was collected on the correspondents during the compilation of the corpus. The STRATAS project currently in progress aims it make some of it available through a dedicated search platform.

But contextualizing a historical corpus does not stop here. In fact this is only the beginning: the ways in which these metadata should be interpreted require familiarity with their socio-historical contexts, ranging from the contemporary society at large down to the physical space the writers inhabited. We have indicated some direc-

tions that this research could take in Nevalainen (2015) and Nevalainen & Raumolin-Brunberg (2017). I will discuss them in my talk.

References

CEEC = Corpus of Early English Correspondence. 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of Modern Languages, University of Helsinki.

Corpora of Early English Correspondence. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html

Nevalainen, T. (2015). What are historical sociolinguistics? Journal of Historical Sociolinguistics 1(2), 243–269. doi:10.1515/jhsl-2015-0014.

Nevalainen, T. & Raumolin-Brunberg, H. (2017). Historical Sociolinguistics: Language Change in Tudor and Stuart England. 2nd edition. London & New York: Routledge.

Parsed Corpus of Early English Correspondence, parsed version. 2006. Annotated by Taylor, A., Nurmi, A., Warner, A., Pintzuk, S. & Nevalainen, T. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

PCEEC: http://www-users.york.ac.uk/~lang22/PCEEC-manual/corpus_description/index.htm

STRATAS = http://www.aka.fi/globalassets/32akatemiaohjelmat/digihum/hanke-esitteet/nevalainen-digihum.pdf

• • •

Irma Taavitsainen
University of Helsinki
irma.taavitsainen@helsinki.fi

Turo Hiltunen
University of Helsinki
turo.hiltunen@helsinki.fi

What linguists need to know about early medical writing?

Early medical writing is a key area for understanding the history of scientific English. At the same time, it is an area that presents many problems for corpus linguists, ranging from obscure terminology and non-standard spelling to broader issues of representativeness and balance in corpus design. The aim of this paper is to introduce and describe some of these issues, and discuss the solutions adopted in the compilation of three diachronic corpora: *Middle English Medical Texts* (MEMT, 1375-1500),

Early Modern English Medical Texts (EMEMT, 1500-1700), and *Late Modern English Medical Texts* (LMEMT 1700-1800). These corpora have been designed to enable the diachronic study of medical writing, and together they contain over 4 million words and cover the full range of linguistic variation over four centuries. Our main focus is on representativeness and use of the data. Given the dramatic changes in language and medicine, as well as societal and cultural changes, how do we assemble a representative text collection? And how do we communicate these decisions to linguists using the corpus, so that they can meaningfully interpret the results that it provides?

Middle English medicine had two textual traditions with a wide scope, from texts of highest learning based on Latin models to practical health guides and remedy books deriving from Old English. MEMT is mainly based on manuscript editions. EMEMT provides continuation to both learned and more popular traditions, and introduces the new categories of writing emerging at the end of the period. In contrast to MEMT, it is based on printed materials LMEMT is nearer to Present-day practices in many respects, but needs sociohistorical contextualizing and much of the material is totally unexplored. Our texts are selected in collaboration with medical historians and our sampling is informed by medical history. For contextualizing the data, information about the textual and communicative traditions is integrated into the corpora, including background metadata with sociolinguistic anchoring of authors and audiences, as well as links to the original text pages.

References

For more information, see Taavitsainen, Hiltunen, I. T., Lehto, A., Marttila, V., Pahta, P., Ratia, M., Suhr, C. & Tyrkkö, J. (2014). Late Modern English Medical Texts 1700–1800: A corpus for analysing eighteenth-century medical English. ICAME Journal 34. Volume 38, Issue 1, 137–153.

• • •

Jacob Thaisen
University of Oslo
jacob.thaisen@ilos.uio.no

Mediaeval texts in their material context

This presentation discusses how linguistic variation present in mediaeval texts may be due to the physical manuscript housing them. It is hardly any new observation that the physical form of a mediaeval text should be an explanatory variable, but it is fair to say that scholars tend to pay lip-service to it rather than give it any serious attention. Material philology is the exception, since this school considers a text's physical form an integral part of its meaning; but linguistics is not this school's primary

concern. Textual scholars sympathetic to it furnish their editions with images of the corresponding manuscript pages. The availability of such editions make it possible to study the variable, whereas corpora catering specifically to the needs of linguists rarely contain any images at all and give sparse bibliographical information about the manuscripts. It is not possible to tell from a concordance if a scribe added <-e> to words to fit a text within a pre-set space. Linguistics corpora may be annotated for syntax but not for the dimensions of the manuscript page. To illustrate the ways in which the physical makeup of a manuscript can shape the selection of variant, the presentation evaluates the distribution of various orthographic variants between the Hengwrt and Ellesmere manuscripts of Chaucer's Canterbury Tales. These variants differ markedly in usage frequency between the two manuscripts, and previous scholarship has attributed this difference to chronological developments within the English of mediaeval London.

• • •

Plenaries

Gaëtanelle Gilquin
Université catholique de Louvain

Hic sunt dracones: Exploring *terra incognita* in learner corpus research

Learner corpus research, by applying the tools and techniques of corpus linguistics to the study of learner language, has made it possible to investigate aspects of interlanguage that had been neglected in traditional second language acquisition research. This has led, for example, to considerable advances in the areas of lexico-grammar (collocations, lexical bundles, etc.) and discourse (connectors, involvement features, etc.), and has also allowed for a more quantitative approach to interlanguage (cf. concepts of under- and overuse). Learner corpus research can now be said to have become a mature field of study, as witnessed among other things by the creation of an association (Learner Corpus Association) and a journal (*International Journal of Learner Corpus Research*) specifically devoted to the subject, and the publication of *The Cambridge Handbook of Learner Corpus Research* (Granger et al. 2015).

In this talk, I would like to explore some areas in learner corpus research that have received comparatively little attention up to now but whose study could lead to interesting developments in the field. I will thus discuss the research trend which consists in approaching the process (rather than the mere product) of writing by relying, e.g., on a corpus including several drafts of the same text (like the Hanken Corpus of Academic Written English for Economics – see Mäkinen & Hiltunen 2016 – and the Malmö University-Chalmers Corpus of Academic Writing as a Process – see Wärnsby et al. 2016) or a corpus representing visible corrections in handwritten texts (as in the Marburg corpus of Intermediate Learner English, see Kreyer 2015: 22-24). I will also discuss the emerging idea that foreign varieties of English, like native and institutionalised second-language varieties, may display diachronic changes and that time of language production should therefore be taken into account in learner corpus studies (cf. the discussion in Laitinen (2016) about English as a Lingua Franca). For these and some other areas that are still largely 'terra incognita' in learner corpus research, I will describe the first explorations, if any, which have been carried out, and I will show how these could be taken further and open up new avenues for research in the study of learner language.

References

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

Kreyer, R. (2015). The Marburg Corpus of Intermediate Learner English (MILE). In M. Callies & S. Götz (Eds.), *Learner Corpora in Language Testing and Assessment* [Studies in Corpus Linguistics 70], 13-34.

Laitinen, M. (2016). Ongoing changes in English modals: On the developments in ELF. In O. Timofeeva, A.-C. Gardner, A. Honkapohja, & S. Chevalier (Eds.),

New Approaches to English Linguistics: Building Bridges [Studies in Language Companion Series 177]. Amsterdam & Philadelphia: John Benjamins, 175-196.
Mäkinen, M. & Hiltunen, T. (2016). Creating a corpus of student writing in economics: Structure and representativeness. In M. J. López-Couso, B. Méndez-Naya, P. Núñez-Pertejo, & I. M. Palacios-Martínez (Eds.), *Corpus Linguistics on the Move: Exploring and Understanding English through Corpora* [Language and Computers: Studies in Digital Linguistics 79]. Leiden & Boston: Brill, 41-58.
Wärnsby, A., Kauppinen, A., Eriksson, A., Wiktorsson, M., Bick, E., & Olsson, L.-J. (2016). Building interdisciplinary bridges. MUCH: The Malmö University-Chalmers Corpus of Academic Writing as a Process. In O. Timofeeva, A.-C. Gardner, A. Honkapohja & S. Chevalier (Eds.), *New Approaches to English Linguistics: Building Bridges* [Studies in Language Companion Series 177]. Amsterdam & Philadelphia: John Benjamins, 197-211.

• • •

Michaela Mahlberg
University of Birmingham

Collocation – New Directions and Opportunities for Interdisciplinarity

The concept of 'collocation' is one of the most fundamental concepts in corpus linguistics. It is often considered in relation to work by Firth (1957) who highlights that habitual co-occurrence patterns are crucial to the meaning of a word. Sinclair et al. (2004: 10) define 'collocation' as "the co-occurrence of two items in a text within a specified environment". Corpus software packages tend to include the retrieval of collocates among their standard functionalities. Although the concept seems to have been around for a long time, collocation still has the potential to move corpus linguistics forward by raising new questions and furthering research across disciplinary boundaries. To make this point, I will consider examples from the CLiC project and the analysis of literary texts (joint work with Peter Stockwell). A new functionality of the CLiC app (http://clic.bham.ac.uk/ Mahlberg et al. 2016) is a KWICgrouping option (cf. O'Donnell 2008) to support the viewing of collocations in context. I will also discuss challenges of comparing collocations across corpora and discourses (joint work with Viola Wiegand and Anthony Hennessey) and indicate potential for future work in this area.

References

Firth, J.R. (1957). *Papers in Linguistics 1934-51.* London: Oxford University Press.
Mahlberg, M., Stockwell, P., de Joode, J., Smith, C., O'Donnell, M. B. (2016).

CLiC Dickens – Novel uses of concordances for the integration of corpus stylistics and cognitive poetics, *Corpora*, 11 (3), 433-463.
http://www.euppublishing.com/doi/pdfplus/10.3366/cor.2016.0102
O'Donnell, M. B. (2008). KWICgrouper – designing a tool for corpus-driven concordance analysis. *International Journal of English Studies,* 8 (1), 107-121.
http://revistas.um.es/ijes/article/view/49121/46991
Sinclair, J., Jones, S. & Daley, R. (2004). *English Collocation Studies. The OSTI Report.* (ed. by R. Krishnamurthy). London: Bloomsbury.

• • •

Michael Stubbs
University of Trier

TEXT, INTERTEXT AND MEANING: with illustrations from Conan Doyle's stories about Sherlock Holmes

The Prague Linguistic Circle, which met at Charles University in the 1920s and 1930s, included Czech linguists (Vilém Mathesius and Josef Vachek), Czech and Austrian-Czech literary scholars (Jan Mukařovský and René Wellek), and Russian émigré linguists (Roman Jakobson and Nikolai Trubetzkoy). In a cultural and political history of the group, Toman (1995) discusses their aims of "redefining the nature of linguistics and literary studies".

The group tackled problems which have still not been solved, including the nature of literary language and the relation of author to text. Two significant quotes are:

"The function of poetic language is the maximum foregrounding of the utterance" (Mukařovský 1932).

"We are not interested in the individual psychology of the novelist, but in his novel as an objective social fact" (Jakobson 1943).

The concept of "intertextuality" suggests an empirical approach to these problems. It is used of cases where units of meaning in one text refer to units of meaning in another text, and implies that the meaning of a text does not depend on (the intention of) the author, but on how readers interpret relations between texts.

The term "intertextuality" was introduced in the 1960s, and in the last fifty years a huge literature has appeared: JSTOR returns over one thousand items with "intertext*" in the title (almost all by literary and cultural theorists, few by linguists). However, this intensive discussion has arguably made only minimal conceptual progress. Since the concept has failed to develop significantly over such a long period, perhaps it should be abandoned in favour of something more worthwhile.

Alternatively, we can argue as follows. First, intertextuality can now be studied by new kinds of corpus data and methods, which were previously not available to literary and cultural scholars. Second, the concept of intertextuality is logically related to other concepts, such as reference, semantic units, paraphrase, and evaluative language. Third, a concept is seen to be significant when it is connected, in a natural way, to a complex of other ideas: if such connections can be made, this often leads to intellectual progress.

The essential empirical question is: Can corpus methods reliably identify intertextual references? The essential conceptual question is: What is the logical relation between intertextuality and meaning?

Conan Doyle's stories about Sherlock Holmes (published 1887 to 1927) provide an ideal corpus for studying intertextuality. They refer to identifiable texts (e.g. earlier detective stories) and allude to contemporary ideas, both scientific (e.g. the value of observational data) and pseudo-scientific (e.g. so-called "criminal anthropology").

The stories were clearly influenced by other fictional and non-fictional texts which characterize the intellectual and social world of the late 1800s. These texts are "objective social facts", but relations between texts change over time, and so therefore does the meaning of the texts. Corpus data on these relations could help to solve problems of language, literature and culture which were posed by the Prague Linguistic Circle over eighty years ago. *Corpus et orbis …*

References

Jakobson, R. (1943). Unpublished lecture. Quoted by Toman 1995: 247-48.

Mukařovský, J. (1932). Standard language and poetic language. In P. Garvin (ed). (1964). A Prague School Reader on Esthetics, Literary Structure and Style. Georgetown UP.

Toman, J. (1995). The Magic of a Common Language: Jakobson, Mathesius, Trubetzkoy, and the Prague Linguistic Circle. MIT Press.

• • •

Wolfgang Teubert
University of Birmingham

The Meaning of Corpus Linguistics

It is content that sets language apart from other immaterial structures. A musical score or a mathematical notation represents indexical signs, but not arbitrary ones. They are devoid of content. What they stand for is not up to negotiation. Language signs, on the other hand, are arbitrary. We can and do argue what a sign means. A language sign does not refer to a discourse-external 'thing'; it refers to other signs. This

is why linguistics cannot be a strict science. There is no true meaning. For meaning is not reducible to statistics or any other formal algorithm. Our quest of meaning has to turn to fuzzy textual evidence, and it is only corpus linguistics that can help us. It has the answer to the hermeneutical enterprise, this greatest of human achievements, initiated by Aristotle in the west and Zhuangzi in the east. It is what makes sense of the social, spiritual and natural world.

For me, the corpus-driven approach to meaning implies that we have to closely listen to what utterances tell us about the meaning of other utterances. The meaning of a singular text, or of one of the recurrent text segments of which it is composed, is the entirety of what has been said about it, of how it has been paraphrased. What we need is a tool picking up candidates of what may be such paraphrases. Another tool that would come in handy would detect the intertextual links tying any new utterance to those earlier utterances to which it is a reaction. Discourse is plurivocal. It speaks in many voices. We need software that shows which voice infects subsequent texts, while leaving other texts untouched. We have to learn to which extent one network of texts, held together by a set of lexical items defining it, isolates itself from other such networks, or undermines them. This is why we have to study the diachronic dimension of discourse. The meaning of a recurrent text segment, of a lexical item, is never stable; it evolves, and it evolves differently in different networks. Statistics are, of course, an indispensible tool to extract promising candidates from the corpus. But statistical findings are not a substitute for meaning. It all depends on our research question what constitutes the meaning of a lexical item. Only humans can sort through the candidates picked up by the program and decide what is relevant or not. The fullness of meaning is never available to us, just as any map is not an accurate representation of the land it shows, but an interpretation of it. When we query the meaning of a text or a recurrent text segment, we aim for an interpretation, an interpretation that will be biased by the categories underlying our research question. Corpus linguistics, as I imagine it, provides the practical foundation to the programme of hermeneutics, i.e. the art of interpretation.

It is we, the discourse participants, and not the lexicographers, who make meaning, and we do it together. By interpreting what has been previously said, by remarking and commenting on it, new ideas will emerge. If they are picked up in subsequent utterances they will have an impact on discourse. Making sense of the world is a collective enterprise, and we all can take part in it. Together we have the power to change the reality confronting us in what we are told. Discourse is, in principle at least, democratic. Every person has a voice. It is, however, a freedom we must take care to guard. The current battle of the internet might easily do away with it. We are increasingly swamped with narratives that aren't ours, and if those in charge don't like our narratives, they call them fake news and edit them out. They want only their reality to confront us, depriving us of our voice. Corpus linguistics, as I envisage it, has what it needs to take stock of such subtle changes of meaning designed to force us to accept their reality, not ours. It is up to us to fight back.

• • •

Ondřej Tichý & Jan Čermák
Faculty of Arts, Charles University

English Diachronic Corpus Linguistic: Prague Themes

Recent years saw a growing popularity of quantitative historical linguistics in Prague, especially application of the methodologies of Digital Humanities. We will introduce the Prague take on diachronic corpus linguistics by focusing on some of these current issues the Department of English Language has been engaged in pursuing.

In one of our early attempts at applying corpus methodology to a Praguian approach, we explored Skalička's and Sgall's theory of morphological typology and focused on the **quantification of typological change** in the history of English. This study, introducing information entropy as a measure of predictability of a system and applying it to the exponents of inflectional categories throughout the history of English, confirmed the major direction of typological change in English, but it also pointed out some intriguing trends in more recent history. Since entropy proved to be well-suited to determine the regularization and simplification of morphological paradigm, we used it again in an attempt to **quantify the standardization of English spelling**, where it likewise proved to be a relatively intuitive measure rather than a mere form : type ratio.

Both the richness of early English morphology and the irregularity of its spelling brought our attention to the paucity of highly annotated (i.e. lemmatized, morphologically tagged) corpus material for Old and Middle English and gave rise to the project of **automatic morphological analysis of Old English** that uses pre-generated dictionaries and rule-based grammar in a manner analogous to the morphological analysis of highly inflected languages, such as Czech, but with a special focus on formal variation. To capture the maximum lexical breadth and the fullest formal variety, Bosworth's and Toller's *Anglo-Saxon Dictionary* was chosen as the source of lexical data for the analyser. That, as the end product of the project, is therefore based on the results of an ongoing **digitization project of the Online Anglo-Saxon Dictionary** (www.bosworthtoller.com) that both supplies the data to, and profits from, the morphological analysis, allowing it to connect with multiple external resources.

The following project of ours thematised the issue of **lexical obsolescence and loss** in historical English. It was first inspired by an interest in the interface between typological changes in the word-formation and lexical mortality on the Old/Middle English threshold, which, due to the nature of preserved data, virtually had to avoid corpus-based methodology. For that reason, we extended the focus of our interest in lexical obsolescence and loss to include the Late Modern English period as well where much larger data available allowed for a more robust quantitative analysis that went hand in hand with further refinement of the method.

Apart from the issue of quantification of language change, we also focused on the visualization of its progress throughout the linguistic community. One of such studies was carried out on the PCEEC, a letter corpus richly annotated with sociolinguistic meta-data that by the nature of the epistolary material it contains is essentially a

graph, or a network, and we employed **network analysis** tools such as Gephi to map lexical innovation and its spread through the language community.

The topics to be presented – though varied and on different levels of linguistic description, spelling, morphology and lexis – all display important facets of systemic interconnection and as such are just different aspects of the same endeavour that has characterised Praguian approach over the past hundred years – keen attentiveness to language structure combined with sustained search for methodological advancement.

• • •

Dedicated Plenary Session

Gerry Nelson & John Kirk
Chinese University of Hong Kong & Dresden University of Technology / Belfast

The International Corpus of English (ICE) Project: The Next 25 Years

As announced at ICAME37 in Hong Kong, during the winter 2016-2017, we have undertaken an extensive and wide-ranging questionnaire-based review of the International Corpus of English with each of the 27 national ICE teams as well as with some other interested parties. The aim of the review has been to take stock of the project and its development over the last 25 years, and to identify the best way forward for the next 25 years.

We completed our review by the late winter and circulated it among respondents. We made numerous recommendations for which we sought approval and agreement. The plenary will be the final stage of this process: it will present to the wider ICAME community the recommendations which will have been agreed within the ICE community.

Our plenary is thus an opportunity to set out the guidelines for second generation ICE corpora (ICE Corpora, Mark II), and to encourage anyone interested in undertaking a second generation ICE corpus to do so.

• • •

Full papers

Bas Aarts
UCL
b.aarts@ucl.ac.uk

Sean Wallis
UCL
s.wallis@ucl.ac.uk

Jill Bowie
UCL
j.bowie@ucl.ac.uk

-Ing clauses in English: structure, usage and recent change

Historically the English gerund is an action noun which could gradually also be used as a verb from the Middle English period onwards (Jack 1988, Fanego 1998, 2004). This process has been called 'verbalization' (Fischer and van der Wurff 2006: 178f.). In Present-Day English the reflexes of this development are verbal nouns and *–ing* clauses, as in (1) and (2):

(1) [The deliberate *sinking* of the ship] was a criminal act
(2) [Deliberately *sinking* the ship] was a criminal act.

In these particular cases the bracketed strings function as subject.

In this paper we study the structure and development of PDE *–ing* clauses, investigating how their frequency and use has changed over recent decades. More specifically, we look at the functions that such clauses can perform within their matrix clauses, using the *Diachronic Corpus of Present-Day English* as our dataset.

Our data show that changes in use are principally found in clauses functioning as subject (as in (2) above), direct object or adverbial in their matrix clauses. We will show that there is a statistically significant trend over recent decades for *–ing* clauses increasingly to have an explicit subject of their own. We hypothesize that this shows that the tendency for *–ing* constructions to become less nominal and more clausal has continued in the recent history of English.

References

Fanego, T. (1996). The gerund in early Modern English: evidence from the Helsinki Corpus. Folia Linguistica Historica XVII/1-2, 97-152.

Fanego, T. (1998). Developments in argument linking in early Modern English gerund phrases. English Language and Linguistics 2, 87-119.

Fischer, O. & van der Wurff, W. (2006). Syntax. In R. Hogg & D. Denison (eds.), A History of the English Language. Cambridge: Cambridge University Press, 109-198.

Jack, G. B. (1988). The origins of the English gerund. NOWELLE 12, 15-75.

• • •

Kathleen Ahrens
The Hong Kong Polytechnic University
kathleen.ahrens@polyu.edu.hk

Huiheng Zeng
The Hong Kong Polytechnic University
huiheng.zeng@connect.polyu.hk

Metaphors of Democracy in the Greater China English Editorial Corpus

Conceptual metaphor research focuses on the conceptual patterns that underlie everyday language use. Using linguistic examples, researchers hypothesize how abstract concepts ('target' domains) use concrete concepts ('source' domains) within a given language. Recent research has focused not only the underlying cognitive conceptualizations, but also the underlying ideology of the speaker and found that speakers of the same language can use very different metaphors when talking about the same concept (Ahrens & Lee, 2009; Ahrens, 2011; Ahrens, 2016a, b; Ahrens & Chan, 2016).

This paper, which is part of the research project on *VARIETIES OF ENGLISH IN THE INDO-PACIFIC (VEIP),* examines the differences in conceptual metaphor use related to the idea of 'democracy' in editorials in contemporary English language newspapers in Hong Kong, China and Taiwan, with a particular focus on how metaphor use reflects the ideology of the editorial columnists in the different regions. We undertake this task by independently compiling an English corpus of editorials from two national newspapers in the People's Republic of China (China Daily and Global Times) during the period of 2011 to 2016 (61,904 words), from three newspapers in Hong Kong (South China Morning Post, The Standard and Hong Kong Free Press) for the period of 2012-2016 (51,216 words), and from two newspapers in Taiwan (The China Post and Taipei Times) for the period of 1999-2016 (33,761 words).

After compiling the corpus, key-word-in-context searches are run using WordSmith tools and downloaded into Excel files, after which the lexeme "democracy" (as well as all the lemmas associated with "democracy" such as "democratic", "democratically", "democrat", "democratize" and "democratization") are analyzed for both tokens andtypes. In this study, metaphorical uses were identified based on Group (2007) guidelines and normalized ratios (NR)are calculated (number of metaphorical instances/corpus size multiplied by 100,000).

Findings to date (based on our analysis of the keyword "democracy" (and all the lemmas of 'democracy') show that PRC discusses this topic significantly less (NR = 8.08) than in the Hong Kong (NR = 29.29) or Taiwan corpus (NR = 38.51). In addition, the Taiwan corpus has two source domains (building and patient) primarily associated with this target domain, reflecting the writers' concerns with the overall robustness of the democratic system in Taiwan, while the Hong Kong corpus has more occurrences of war and journey metaphors, indicating writers' concerns with protecting democratic rights as well as questions about the how far Hong Kong can "go" in terms

of democracy as a Special Administrative Region of China. We plan on adding to this analysis by looking at other keywords related to democracy, including words that relate to democratic practices such as elections, voting, balloting, and campaigning to gather a fuller picture of the viewpoints of the writers with relation to these concepts.

Given that each of these regions practices democracy to a greater or lesser degree, compiling and analyzing the data will allow better understanding of how opinion influencers in the media utilize conceptual metaphors to sway theirreaders. In addition, study has particular significance with respect to Hong Kong as it navigates demands for an increase level of participation in democratic processes. By better understanding the way issues relating to democracy are conceptualized throughout greater China, the more likely it is the various stakeholders and concerned citizens can understand others' viewpoints and communicate effectively.

References

Ahrens, K. (2011). Examining conceptual metaphor models through lexical frequency patterns: A case study of U.S. presidential speeches. In H. J. Schmid (Ed.), Windows to the mind. Series: Applications of cognitive linguistics (pp. 167-184). Berlin: Mouton De Gruyter.

Ahrens, K. (2016a). English conceptualizations in a Chinese context: BUILDING metaphors in Hong Kong policy addresses. Presented at International Conference on Boundary Crossing and Bridge Building. National Taiwan University, Taipei, Taiwan, 17 June, 2016.

Ahrens, K. (2016b). Mappings from the source domain of war in Hong Kong policy addresses. Presented at Metaphor Festival Conference. University of Amsterdam, Amsterdam, 31 August - 3 September, 2016.

Ahrens, K. & Chan, I. W. S. (2016). The role of mapping principles in the translation of political speeches. Presented at First International Conference of Intercultural Linguistics, Prato, Italy, 20-22 July, 2016.

Ahrens, K. & Lee, S. Y. M. (2009). Gender versus politics: When conceptual models collide in the U.S. Senate. In K. Ahrens (Ed.), Politics, gender, and conceptual metaphors (pp. 62-82). Basingstoke and New York: Palgrave-MacMillan.

• • •

Karin Aijmer
University of Gothenburg
karin.aijmer@eng.gu.se

The intensifier *fucking* in Spoken BNC14

Intensifiers hold a great fascination for scholars because of their variability and the changes they undergo. Once they get a foothold in society and achieve general acceptance they become subject to delexicalization and may be replaced by other intensifiers. Such developments have been studied diachronically but can also be described in a shorter micro-diachronic perspective using synchronic cross-generational data.

The purpose of this study is to identify the syntactic, semantic, pragmatic and sociolinguistic factors accounting for the variability of *fucking* and the changes it has undergone over a short period of time. The presentation takes advantage of the possibility to use the 'new' British National corpus (SpokenBNC2014Early Access Subset) representing data from the later part of the 20th century. BNC2014 contains approximately five million words of spoken English involving 376 speakers. The data consists of conversations on a variety of different topics between two or more speakers who are friends or family members. Information about the speakers (age, gender, social class, education, regional dialect) has been coded in the corpus. The intensifier *fucking* is of particular interest since it can be shown to have more than doubled its frequency in BNC14 compared with the older BNC.

Intensifiers can be analysed along several parameters such as type and degree of evaluation or expressivity (cf Cacchiani 2005), and the semantic type of adjective they co-occur with. The most frequent adjectives are evaluative (*good, happy, bad)* but non-evaluative adjectives are also found (*expensive).*The adjectives (with or without intensifiers) are typically used to signal the speaker's subjective evaluation associated with 'affect', judgement of personal traits or appreciation and emotional reaction (Martin 2000). The patterns generally contain *be* ('it/that is INT+ADJ'*)* but patterns with attributive adjectives ('Intensifier +Adjective +NP') are also found. Intensifiers can turn into stronger emotional expressions where they combine with 'extreme' (or 'superlative') adjectives (Cacchiani 2005:410) occupying a position at the positive or negative end of a scale. Extreme adjectives include trendy ones such as *brilliant, amazing, weird* which are typically used by adolescents to express a certain style or 'persona'.

References

Cacchiani, S. (2005). Local vehicles for intensification and involvement: the case of English intensifiers. In P. Cap, (ed.), *Pragmatics Today*. Frankfurt am Main: Peter Lang. 401-419

Martin, J. (2000). Beyond exchange: APPRAISAL systems in English. In S. Hunston & G. Thompson (eds), *Evaluation in Text. Authorial Stance and the Construction of Discourse.* Oxford: OUP. 142-175.

• • •

Udo Baumann
University of Freiburg
udo.baumann@frequenz.uni-freiburg.de

Caught between convention and innovation – The case of the progressive in spoken British English

The recent frequency increase of the English progressive construction has been addressed by numerous studies. Mair and Hundt (1995) were the first to show that it was still under way in late 20th-century English. Subsequent studies (e.g. Leech, Hundt, Mair & Smith 2009) have argued that the progressive's increasing use has been mainly taking place within the established formal and functional contexts (i.e. aspectual uses of the present progressive active). Non-prototypical uses with stative verbs are regarded as a statistically marginal phenomenon. Contrary to this view, Levin (2013) shows that progressive use with stative verbs has increased considerably in 20th-century AmE, identifying it as a contributing factor to the construction's overall increase.

This paper moves the focus to 21st-century spoken BrE, asking whether the progressive's frequency development is still under way – thus, illuminating the current state of the phenomenon. Furthermore, it is asked whether the progressive's frequency increase is mainly occurring within the boundaries of its conventionalized contexts of use or rather connected to a spread to new formal and verbal contexts – a question that has so far produced seemingly contradictory results.

Specifically, I analyzed progressive use in the Diachronic Corpus of Present-day Spoken English (DCPSE), comprising data from the 1950-70s and from the early 1990s. Furthermore, I used a new, self-compiled corpus of spoken BrE, containing data from 2012-16, designed to match the DCPSE. The data comprises 1.2 million words, which is justifiable for the analysis of a high frequency construction as the progressive. To uncover changes in the progressive's morphosyntactic and semantic profile, different frequency metrics as well as Distinctive Collexeme Analysis (Gries & Stefanowitsch 2004; Hilpert 2012) and Hierarchical Configural Frequency Analysis (Gries 2004; Hilpert 2013) were used.

The results first reveal that the progressive's frequency increase continues in all but one of the analyzed genres. While frequencies no longer increase in Face-to-Face Conversation, Broadcast Interview & Discussion, Spontaneous Commentary, and Parliamentary & Prepared Speech show a significant upward trend. Second, in all genres that exhibit increasing progressive use, it is the present progressive active that accounts for most of the increase. Regarding different verb classes, it is progressives with activity/event verbs that have increased most in frequency. Prototypical combinations of the present progressive active with activity/event verbs have become even more frequent over time. However, supposedly stative verb classes (existence/relationship and perception/sensation) also account for roughly a quarter of the progressive's frequency increase in all genres. *Have* and *feel* are among the verbs whose use with the progressive has increased most.

In sum, the results suggest that the progressive's frequency increase still has not subsided yet and that it has spread from conversation to more conservative spoken genres. While most of the increase seems to be happening within the conventionalized contexts of use, part of the increase is due to a higher readiness to use the construction with stative verbs. This is in line with De Smet's (2016) model of change, which proposes that unconventional uses become more likely as the related conventional coding solution becomes more frequent and mentally accessible.

References

De Smet, H. (2016). How gradual change progresses: The interaction between convention and innovation. Language Variation and Change, 28, 83-102.

Gries, S. T. (2004). HCFA 3.2. A program for R for Windwos 2.X [Software].

Gries, S. T. & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. International Journal of Corpus Linguistics, 9(1), 97-129.

Hilpert, M. (2012). Diachronic collostructional analysis: How to use it and how to deal with confounding factors. In K. Allan & J. A. Robinson (Eds.), Current methods in historical semantics (pp. 133-160). Berlin, Germany: De Gruyter.

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

Levin, M. (2013). The progressive verb in modern American English. In B. Aarts, J. Close, G. Leech & S. Wallis (Eds.), The verb phrase in English: Investigating recent language change with corpora (pp.187-216). Cambridge, England: Cambridge University Press.

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

Mair, C. & Hundt, M. (1995). Why is the progressive becoming more frequent in English? A corpus-based investigation of language change in progress. In W. Riehle & H. Keiper (Eds.), Anglistentag 1994 Graz (pp.247-254). Tübingen, Germany: Max Niemeyer.

• • •

Eva Berlage
Universität Hamburg
Eva.Berlage@uni-hamburg.de

The semantic-syntactic specialization of composite predicate constructions

In this paper I want to explore the semantic-syntactic evolution of a series of composite predicate constructions (from now on: CPCs) in English from the 16th century to the present-day. Examples include such constructions as *make fun of*, *make mention of* and *take delight in*. My aim is, firstly, to place them on a cline that ranges from constructions that are completely syntactically fixed to those that are syntactically flexible (e.g. allow for modifiers and determiners to intervene between the verb and the complement as in *make good use of* and *make no mention of*). Secondly, I want to predict which constructions undergo semantic-syntactic specialization over time and attempt to explain this phenomenon.

To date, the evolution of CPCs has mainly been explained from the perspective of the verb or the complement involved (see e.g. Brinton & Traugott 2005; Brinton 2008; Trousdale 2008; Berlage 2012). Only a single pilot study (Berlage 2014) focusing on *take advantage of* and *make use of* has addressed the question of whether CPCs that have morphologically related simple verb counterparts (e.g. *make use of – use*) show a development different from those without such counterparts (e.g. *take advantage of*). My paper takes up this idea, linking synchronic variation to change over time.

The present study subjects a series of CPCs to an investigation in terms of the 'noun phrasiness' of the complement. Referring back to Ross (1973; 1995; for a semantic approach, see Nunberg et al. 1994), I will measure the degree to which the complements take modifiers and/or determiners, indicating how autonomous the noun phrase in the CPC is or, inversely, how fixed the CPC has become.

Corpus analyses based on more than 100 million words of historical British fiction reveal that there are many CPCs for which we observe either a strong increase in or very high ratios of determiners throughout (as e.g. in *take no notice of*). Closer inspection shows that these determiners mainly occur in negative or other non-assertive contexts like interrogatives (*Are you taking any notice of them?*) and comparative-like constructions (*[…] was too busy eating to take much notice of him*). For some constructions, the association with these non-assertive contexts is so strong that they gradually turn into what we call 'negative polarity items' (Huddleston/Pullum et al. 2002: 823), being barred from assertive contexts.

Comparing those constructions that become semantically and syntactically specialized to those that do not, we see that specialization only occurs where there is a simple (or prepositional) verb alternant in the paradigm (as in the cases of *make mention of – mention, take delight in – delight in*). My paper argues that specialization requires the existence of an alternative item that can take over the remaining functions.

References

Berlage, E. (2012). At the interface of grammaticalisation and lexicalisation: the case of take prisoner. English Language and Linguistics 16, 35-55.

Berlage, E. (2014). Opposite developments in composite predicate constructions: the case of take advantage of and make use of. In M. Hundt (ed.), Late Modern English Syntax. Cambridge: CUP, 207-223.

Brinton, L. (2008). Where grammar and lexis meet: composite predicates in English. In E. Seoane & M. J. López-Couso (eds.), Theoretical and Empirical Issues in Grammaticalization. Amsterdam/Philadelphia: Benjamins, 33-53.

Brinton, L. J. & E. C. Traugott (2005). Lexicalization and Language Change. Cambridge: CUP.

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

Nunberg, G., Sag, I. A. & Wasow, T. (1994). Idioms. Language 70, 491-538.

Ross, J. R. (1973). A fake NP squish. In C. J. N. Bailey & R. W. Shuy (eds.), New Ways of Analyzing Variation in English. Washington, D.C.: Georgetown University Press, 96-140.

Ross, J. R. (1995). Defective noun phrases. In A. Dainora, R. Hemphill, B. Lukas, B. Need & S. Pargman (eds.), Proceedings of the Thirty- first Regional Meeting of the Chicago Linguistic Society. University of Chicago: Chicago Linguistic Society, 398-440.

Trousdale, G. (2008). Constructions in grammaticalization and lexicalization: evidence from the history of composite predicate constructions in English. In G. Trousdale & N. Gisborne (eds.), Constructional Approaches to English Grammar. Berlin: Mouton de Gruyter, 33-67.

• • •

Zeltia Blanco-Suárez
University of Cantabria
zeltia.blanco@unican.es

Mario Serrano-Losada
University of Santiago de Compostela
mario.serrano@usc.es

***Needless to say* and *goes without saying*: Diachronic and synchronic considerations**

Speakers/writers often resort to certain strategies to convey the obviousness of their statements, hence acknowledging that the information expressed is known, expected or self-evident. Consider in this regard the expressions *needless to say* and *goes without saying* in (1)-(2) below:

(1) *The whole sea seemed alive that night and there was little sleep for the crew*, **needless to say**. (BYU-BNC, 1989)
(2) *No need to tell you how sorry I am*—**goes without saying**—*but I wouldn't call that murder, would you? Unless you're accusing Pa.* (BYU-BNC, 1993)

Such forms are used both as hedging devices to shade categorical assertions and for face-saving purposes. They also serve as evidential strategies (see Chafe 1986, Aikhenvald 2004, Cornillie 2009, Cruschina & Remberger 2008, Diewald & Smirnova 2010) whereby the speaker/writer provides evidence for what s/he is saying, thus emphasizing a shared worldview or experience and drawing attention to aspects of general knowledge. Additionally, such expressions have an interactional and intersubjective function (see López-Couso 2010, Traugott 2003, 2010, 2012), aimed at seeking agreement with the interlocutor, thus conveying "the attitude of the speaker towards the content of the utterance and/or the degree of speaker endorsement" (Dehé & Kavalova 2007: 1).

The aim of the present research is twofold. On the one hand, it sets to trace the diachronic development of the PDE evidential parentheticals *needless to say* and (*it*) *goes without saying*. On the other, it delves on the frequency and distribution of both expressions in Present-day British and American English, examining the degree of prevalence of each of these forms in the different textual genres in order to determine potential patterns of distribution.

Our preliminary results show that parenthetical *needless to say* has its roots in the EModE <*needless* + extraposed *to*-INF SBJ> construction (meaning 'it is unnecessary to do something'), which originally licensed a wide range of infinitives. In the course of time, the construction became restricted to uses with utterance predicates, eventually giving rise to the fixed evidential parenthetical with the verb *say*. Although the construction did not encode evidential nuances at first, pragmatic inferences of an evidential nature became conventionalized during the LModE period ('that which is

unnecessary to say is obvious'), when the first parenthetical uses of the construction are attested. In contrast to *needless to say*, which dates to the sixteenth century, (*it*) *goes without saying* is only attested from the nineteenth century on, being a borrowing from the much older French expression *cela va sans dire*. Moreover, while in PDE parenthetical *needless to say* occurs predominantly at left periphery with forward scope, (*it*) *goes without saying* is mostly attested in the right periphery, in utterance final position. Data for the present paper have been drawn from several diachronic and synchronic sources, including the OED, EEBOCorp, CLMET, Hansard Corpus, BYU-BNC, COHA and COCA.

Sources

BYU-BNC = *Brigham Young University-British National Corpus*, compiled by M.Davies. 2004-. <http://corpus.byu.edu/bnc/>.

CLMET = *The Corpus of Late Modern English Texts*, *version 3.0*, compiled by H. De Smet, H. Jürgen Dillerand & J. Tyrkkö. <https://perswww.kuleuven.be/~u0044428/>.

COCA = *The Corpus of Contemporary American English*, compiled by M. Davies. 2008-. <http://corpus.byu.edu/coca/>.

COHA = *The Corpus of Historical American English*, compiled by M. Davies. 2010-. <http://corpus.byu.edu/coha/>.

EEBOCorp = *Early English Books Online Corpus* 1.0, compiled by P. Petré. 2013. <https://lirias.kuleuven.be/handle/123456789/416330>.

Hansard Corpus = *The Hansard Corpus 1803-2005*, compiled by M. Alexander & M. Davies. 2015-. <http://www.hansard-corpus.org>.

OED = *Oxford English Dictionary Online*. Oxford University Press. <http://www.oed.com>.

References

Aikhenvald, A. (2004). Evidentiality. Oxford: Oxford University Press.

Chafe, W. (1986). Evidentiality in English conversation and academic writing. In W. Chafe & J. Nichols (Eds.), Evidentiality: The linguistic coding of epistemology. Norwood, NJ: Ablex , 261-272.

Cornillie, B. (2009). Evidentiality and epistemic modality: On the close relationship between two different categories. Functions of Language, 16 (1), 44-62.

Cruschina, S. & Remberger, E. M.. (2008). Hearsay and reported speech: Evidentiality in romance. Rivista di Grammatica Generativa, 33, 95-116.

Dehé, N. & Kavalova, Y. (Eds.). (2007). Parentheticals. Amsterdam: John Benjamins.

Diewald, G. & Smirnova, E. (2010). Linguistic realization of evidentiality in European languages. Berlin: Mouton de Gruyter.

López-Couso, M. J. (2010). Subjectification and intersubjectification. In A. H. Jucker & I Taavitsainen (Eds.), Historical pragmatics. Berlin: Mouton de Gruyter, 127-163.

Traugott, E. C. (2003). From subjectification to intersubjectification. R. Hickey (Ed.), Motives for Language Change. Cambridge: Cambridge University Press, 124-139.

Traugott, E. C. (2010). (Inter)subjectivity and (inter)subjectification: A reassessment. In K. Davidse, L. Vandelanotte & H. Cuyckens (Eds.), Subjectification, intersubjectification and grammaticalization.Berlin: Mouton de Gruyter, 29-70.
Traugott, E. C. (2012). Intersubjectification and clause periphery. English Text Construction 5 (1), 7-28.

• • •

Tine Breban
The University of Manchester
tine.breban@manchester.ac.uk

Operationalizing semantic change: a micro-analytical approach to identifying grammatical and lexical uses of certain

This paper is concerned with the issue of operationalizing changes primarily defined on the basis of meaning. Through operationalization the main strategy used in corpus studies of providing a semantic analysis of individual examples in context is sought to be made more objective and efficient. Operationalization is a hot topic, see e.g. recent papers on operationalization of (inter)subjectification (a.o. Torres Cacoullos & Schwenter 2005; López-Couso 2010; Traugott 2010, 2015; Brems et al. 2014). This paper revisits a type of semantic change, for which the question of operationalization is often considered to have been answered some time ago, grammaticalization. In many corpus studies, Lehmann's (1982) parameters are applied to identify and justify a grammaticalization analysis. However, the parameters have been argued to be found only in advanced grammaticalization (Hopper 1991) and to be tailored to synthetic languages. In addition, as pointed out by Norde (2012), researchers often cherry-pick rather than systematically apply parameters. The aim of this paper is to suggest supplementary means for operationalization which are more suitable for English. The theoretical backdrop is the idea that meaning is coded in lexico-grammatical form (e.g. Halliday 1994). Crucially, the concept of 'form' is broader than e.g. in Lehmann, and also includes distributional behaviour (Author 2014). Concretely, what I will look for in corpus data is evidence for distributional behaviour associated with a lexical or a grammatical meaning, and changes in frequency of attestations of this behaviour count as evidence for semantic change.

I will put this type of micro-level operationalization of grammaticalization into practice using a case study involving the adjective *certain*, which has a range of lexical (e.g. 'sure' in a certain victory) and grammatical meanings (e.g. as indefinite (post)determiner/quantifier in a *certain girl*). The first step is to identify distributional behaviours associated with the different meanings based on dictionary entries and existing studies of *certain*, as well as of behaviour associated with the different functions in

general. The choice of *certain* is motivated by the extensive (corpus-based) literature on different functions of adjectives in the noun phrase. Examples of behaviour include specific collocating nouns for the lexical meanings, differing relative positions with regard to other adjectives in the noun phrase, occurrence in syntactic constructions, e.g. presentational constructions, which are a functional fit for the grammatical meanings. The second step is to apply theses distributional profiles to historical data from the Penn family of corpora; a key focus is efficiency of operationalizability, using POS-tagging and parsing where possible.

I end the paper by evaluating the process. The pros are that this kind of operationalization avoids assigning meaning to individual historical examples, and can be sped up by making use of tagging and parsing. The cons are that it does require detailed micro-level analysis to start with. However, this analysis can be transferred from/to other items with similar functions. The more general conclusion is that distributional behaviour is 'local' (e.g. tied to the (English) noun phrase). This suggests that the level of granularity to operationalize semantic change is the local micro-level.

References

Brems, L., Ghesquière, L. & Van de Velde, F. (Eds.). (2014 [2012]). Intersubjectivity and intersubjectification in grammar and discourse. Amsterdam & Philadelphia: John Benjamins.

Halliday, M. A. K. (1994). An Introduction to Functional Grammar. 2nd edn. London: Arnold.

Hopper, P. J. (1991). On some principles of grammaticization. In E. C. Traugott & B. Heine (Eds.), Approaches to grammaticalization. Vol 1. Amsterdam & Philadelphia: John Benjamins, 17-35.

Lehmann, C. (1982). Thoughts on grammaticalization: A programmatic sketch. Arbeiten des Kölner Universalien-Projektes. University of Cologne, Institut für Sprachwissenschaft. [Reprinted as (1995.) Thoughts on grammaticalization. München: LINCOM Europe].

López-Couso, M. J. (2010). Subjectification and intersubjectification. In A. Jucker & I. Taavitsainen (Eds.), Historical pragmatics. Berlin: Mouton de Gruyter, 127-163.

Norde, M. (2012). Lehmann's parameters revisited. In K. Davidse, T. Breban, L. Brems & T. Mortelmans (Eds.), Grammaticalization and language change: New reflections. Amsterdam & Philadelphia: John Benjamins, 73-110.

Torres Cacoullos, R. & Schwenter, S. (2005). Towards an operational notion of subjectification. In R. T. Cover & Y. Kim (Eds.), Proceedings of the 31st Annual Meeting of the Berkeley Linguistics Society. Berkeley, CA: Berkeley Linguistics Society, 347-358.

Traugott, E. C. (2010). (Inter)subjectivity and (inter)subjectification: A reassessment. In K. Davidse, L. Vandelanotte & H. Cuyckens (Eds.), Subjectification, intersubjectification and grammaticalization. Berlin: Mouton de Gruyter, 29-69.

Traugott, E. C. (2015). Identifying micro-changes in a particular linguistic change-type: the case of subjectification. In M. Kytö & P. Pahta (Eds.), The Cambridge handbook of English historical linguistics. Cambridge: Cambridge University Press, 351-375.

• • •

Marie-Louise Brunner
Trier University of Applied Sciences
ml.brunner@umwelt-campus.de

Stefan Diemer
Trier University of Applied Sciences
s.diemer@umwelt-campus.de

Selina Schmidt
Birmingham City University
selina.schmidt@mail.bcu.ac.uk

Multimodal meaning making: Developing a taxonomy for the transcription of gesture in a corpus of Skype conversations

Gestures have been increasingly studied as a key means of meaning making (cf. e.g. McNeill 2000, Streeck 2010), and there have been calls for a stronger consideration of multimodal elements in corpora (e.g. Adolphs & Carter 2013). Quantitative gesture analyses are rare, not least due to a lack of multimodal corpora that allow for the detailed study of non-verbal aspects in corpus linguistics. One of the main issues with quantifying gestures in a corpus is their retrievability, if they are transcribed at all, as they are difficult to systematize. Such a systematic approach is, however, indispensable to ensure a consistent and "replicable coding scheme" (see Adolphs and Carter 2013:155) which allows, in turn, to quantify results.

We use CASE, the Corpus of Academic Spoken English (forthcoming) as a basis for developing a tentative taxonomy of gestures. CASE consists of Skype conversations between speakers of English as a Lingua Franca from eight European countries. For quantitative analysis, we use a subcorpus of 20 conversations, BabyCASE (forthcoming). BabyCASE consists of 13 hours of Skype conversations, totaling roughly 115 000 words in the annotated version. The interaction between verbal discourse and non-verbal elements in CASE allows a differentiated view that has not yet been explored in other corpora.

We follow a bottom-up approach in developing a taxonomy for gestures in CASE. Gestures contributing to meaning making were marked in a descriptive way by transcribers and then extracted and grouped in systematically retrievable descriptive categories. Our taxonomy is illustrated with a case study of the eight most frequent gestures in our data: nods, head shakes, shrugs, pointing, air quotes, imitating gestures, waving, and physical stance shifts.

Gestures cannot be considered in isolation, but as being interconnected with verbal interaction in a dynamic process of meaning making. Quantitative and qualitative methods were employed to analyze how gestures contribute to the negotiation of meaning in interaction. Keyword and context analyses were used to isolate co-occurring items, allowing additional categorization and quantification. Nodding, for example, most frequently co-occurs with *yeah*, *mhm*, *right*, and *okay*, indicating (and

emphasizing) support and agreement. However, such co-occurrences could only be observed for two thirds of the cases, leaving the remaining instances open for interpretation; likewise, many of the other gestures had less obvious or frequent co-occurrences. Those instances of gestures were qualitatively analyzed to further categorize the different levels of meaning. Headshakes, for example, can have multiple, and even opposing meanings (e.g. confirmation and negation, awe and despair, resignation, lack of understanding, etc.), depending on conversational context and speaker background (see also Brunner, Diemer and Schmidt forthcoming).

Our findings suggest that a descriptive categorization is essential when creating a taxonomy of gestures suitable for corpus analysis, and that a mixed-methods quantitative and qualitative approach allows for a more nuanced and complete interpretation. Our paper thus contributes to the integration of rich, multimodal data as part of the analysis of spoken language corpora.

References

Adolphs, S. & Carter, R. (2013). Spoken corpus linguistics: From monomodal to multimodal [Routledge advances in corpus linguistics, 15]. Routledge.

BabyCASE. (Forthcoming). Corpus of Academic Spoken English – 20 conversation subcorpus. Saarbrücken: Saarland University and Birkenfeld: Trier University of Applied Sciences, Environmental Campus Birkenfeld.

Brunner, M.-L., Diemer, S., & Schmidt, S. (Forthcoming). "… okay so good luck with that ((laughing))?" – Managing rich data in a corpus of Skype conversations. Studies in Variation, Contacts and Change in English. Helsinki: Varieng.

CASE. (Forthcoming). Corpus of Academic Spoken English. Saarbrücken: Saarland University and Birkenfeld: Trier University of Applied Sciences, Environmental Campus Birkenfeld. [http://www.uni-saarland.de/index.php?id=48492] (03.04.2016).

McNeill, D. (2000). Language and gesture [Language, culture, and cognition, 2]. Cambridge University Press.

Streeck, J. (2010). Gesturecraft: The Manu-facture of Meaning. [Gesture studies, 2]. Benjamins.

• • •

Beatrix Busse
Heidelberg University
beatrix.busse@as.uni-heidelberg.de

Kirsten Gather
Heidelberg University
kirsten.gather@uni-heidelberg.de

Ingo Kleiber
Heidelberg University
kleiber@heiedu.uni-heidelberg.de

HeidelGram: A corpus-based network analysis of grammarians' references in 19th-century British grammars

The *HeidelGram* project, based at the English Department of Heidelberg University, has a twofold aim. It makes an essential contribution to historical grammar studies by compiling, investigating, and making available a representative 10-million-word corpus of historical English grammar books from the 16th to the 19th centuries, and it introduces state-of-the-art network analysis into diachronic corpus linguistics in order to considerably extend the set of concepts and methods applied in historical linguistics and corpus linguistics, and to exemplarily implement and analyse various kinds of networks, such as a network of grammarians, and a network of manifestations of language purism, such as *verbal hygiene* (Cameron 1995), in long-term diachrony.

In contrast to social network analyses of historical material (e.g. Bergs 2005, Sairio 2009, Fitzmaurice 2010) and to network studies based on fictional texts (e.g. Agarwal et al. 2012, Moretti 2013), the combination of corpus-based diachronic linguistics and network analysis is rather uncharted territory. This pilot project constitutes the first part of a series of diachronic network analyses of historical English grammar books.

The present study investigates the relationships between 19th-century grammarians by examining references authors make to other grammars and grammarians. Based on White's notion of 'scholarly networks', references are understood as „record[s] of who has cited whom within a fixed set of authors" (White 2011: 275) irrespective of their personal acquaintance.

A pilot corpus of 19th-century British grammar books (40 texts, ca. 2.6 mio. words) forms the basis for this kind of network analysis. It contains the most well-known and widely distributed grammars of the 19th century (cf. Leitner 1986, 1991, Linn 2006, Michael 1987, Görlach 1998) as full texts in digitised form. Main criteria for text selection are numbers of editions, distribution, and common use of grammars, as found in book catalogues and secondary literature on grammar writing.

A list of English and foreign grammarians from the 16th to 19th centuries that are nowadays usually considered the most famous and influential authors of their time (cf. Dons 2004, Finegan 1998, Görlach 1998, Linn 2006, Schmitter 1996, Tieken-Boon

van Ostade 2008, Wolf 2011) comprises the search terms which, applied to the pilot corpus, yield all references made to other grammarians.

The ties between authors will be examined quantitatively, i.e. in terms of the number of references, and qualitatively, i.e. by classifying different kinds of references, e.g. quotation, approval of approaches to grammar, the citing of authorities, and various forms of criticism. Approval, for instance, is „I *concur* with Baker in considering …" (Crombie (1802) on Baker (1724)), whereas an example of criticism is „Mr. Cobbett has *mistaken* the real causes of defective arrangement" (Doherty (1841) on Cobbett (1818)). We will show different and changing attitudes towards other grammarians, and discuss substantial implications for the development of the genre.

The network of references will further reveal paradigm shifts in grammar writing, indicating particularly the rise of descriptive grammars after the predominance of prescriptivism and critically reflecting on what is often called ‚prescriptive' and ‚descriptive'.

References

Agarwal, A., Corvalan, A., Jensen, J. & Rambow, O. (2012). Social network analysis of Alice in Wonderland. Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature (Association for Computational Linguistics, Montreal, Canada, 2012), 88–96, URL: http://www.aclweb.org/anthology/W12-2513.

Baker, W. (1724). Rules for True Spelling and Writing English. 2nd Edition. Bristol: Joseph Penn.

Bergs, A. (2005). Social Networks and Historical Sociolinguistics. Studies in Morphosyntactic Variation in the Paston Letters (1421-1503). Berlin/New York: Mouton de Gruyter.

Cameron, D. (1995). Verbal Hygiene. The Politics of Language. London: Routledge.

Cobbett, W. (1818). Grammar of the English Language, in a Series of Letters. New York: Clayton and Kingsland.

Crombie, A. (1802). The Etymology and Syntax of the English Language, Explained and Illustrated. London: J. Johnson.

Doherty, H. (1841). An Introduction to English Grammar on Universal Principles. London: Marshall & Co.

Dons, U. (2004). Descriptive Adequacy of Early Modern English Grammars. Berlin/New York: Mouton de Gruyter.

Finegan, E. (1998). English Grammar and Usage. In S. Romaine (Ed.), The Cambridge History of the English Language Vol. IV: 1776-1997. Cambridge: Cambridge University Press, 536-588.

Fitzmaurice, S. (2010). Coalitions, networks, and discourse communities in Augustan England: The Spectator and the early eighteenth-century essay. In R. Hickey (Ed.), Eighteenth-Century English. Cambridge: Cambridge University Press, 106-132.

Görlach, M. (1998). An Annotated Bibliography of Nineteenth-Century Grammars of English. Amsterdam/Philadelphia: John Benjamins.

Leitner, G. (1986). English Traditional Grammars in the Nineteenth Century. In D. Kastovsky & A. Szwedek (Eds.), Linguistics Across Historical and Geographical Boundaries; 2: Descriptive, Contrastive, and Applied Linguistics. Trends in Linguistics. Studies and Monographs [TiLSM] 32. Berlin et al.: Mouton de Gruyter, 1333-55.
Leitner, G. (1991). English Traditional Grammars: An International Perspective. Studies in the History of the Language Sciences 62. Amsterdam: John Benjamins, 1991.
Linn, A. (2006). English Grammar Writing. In B. Aarts & A. McMahon (Eds.), Handbook of English Linguistics. Malden, Mass.: Blackwell, 72-92.
Michael, I. (1987). The Teaching of English. Cambridge: Cambridge University Press.
Moretti, F. (2013). Distant Reading. London: Verso.
Sairio, A. (2009). Language and Letters of the Bluestocking Network. Sociolinguistic Issues in Eighteenth-century Epistolary English. Helsinki: Société Néophilologique (Mémoires de la Société Néophilologique de Helsinki).
Schmitter, P. (Ed.) (1996). Sprachtheorien der Neuzeit. 2 Vols. Tübingen: Gunter Narr Verlag.
Tieken-Boon van Ostade, I. (Ed.) (2008). Grammars, Grammarians and Grammar-Writing in Eighteenth-century England. Berlin/New York: Mouton de Gruyter.
White, H. D. (2011). Scientific and Scholarly Networks. In J. Scott & P.J. Carrington (Eds.), The Sage Handbook of Social Network Analysis, 271-285.
Wolf, G. (2011). Englische Grammatikschreibung 1600-1900 – der Wandel einer Diskurstradition. Arbeiten zur Sprachanalyse 54. Frankfurt a. M.: Peter Lang.

• • •

Marcus Callies
University of Bremen
callies@uni-bremen.de

Patterns of lexico-grammatical variability and innovation in idioms across varieties of English

Research in Cognitive Sociolinguistics (e.g. Geeraerts et al. 2010) and Cultural Linguistics (Sharifian 2015) has highlighted the importance of cultural background knowledge and underlying cultural conceptualizations for the interpretation of lexis and phraseology in varieties of English (see also Wolf & Polzenhagen 2009). In this context, this paper discusses the culture-specificity of the use of figurative language in varieties of English. Idioms as a special type of figurative language are understood as being conceptually motivated by underlying metaphorical mappings, also reflecting the nexus of language and culture (e.g. Gibbs 2007, Kövecses 2005). Generally speak-

ing, when compared to the field of lexico-grammar, there is relatively little research on idiomatic phraseology and figurative language use in varieties of English (see e.g. Platt et al. (1984: 107-110) for a brief overview and numerous examples of the development of idioms in several new Englishes, and the work by Skandera (2003) and others on African Englishes). However, Schneider (2007: 46; 86) highlights the importance of idioms for the description and study of patterns of nativization in new Englishes, arguing that collocational preferences and idiomatic phraseology are extremely characteristic of new Englishes and indicate structural nativization on the lexical level.

On the basis of large web corpora of varieties of English (e.g. Davies 2013), the paper examines the lexico-grammatical and conceptual variability of selected idiomatic expressions related to the source domains FOOD and EATING to answer two research questions: 1) Is there evidence for the lexico-grammatical and conceptual variability of certain idioms across varieties of English? 2) Can (emerging) idioms be taken as culture- and variety-specific "linguistic markers" of certain new Englishes? The results show patterns of lexico-grammatical variation and innovation of idioms in (West) African Englishes and confirm previous research that points towards the high salience and frequency of food and related concepts of eating as source domains in idioms and, more generally, conceptual metaphorical mappings in West African cultures (Wolf and Polzenhagen 2007a, 2007b). The paper concludes that food and eating as source domains seem fruitful points of departure for further studies on culture- and variety-specific "linguistic markers" across varieties of English.

References

Davies, M. (2013). Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries. <http://corpus.byu.edu/glowbe>

Geeraerts, D., Kristiansen, G. & Peirsman, Y. (eds.) (2010). Advances in Cognitive Sociolinguistics. Berlin: de Gruyter Mouton.

Gibbs, R. W. Jr. (2007). Idioms and formulaic language. In D. Geeraerts & H. Cuyckens (Eds.), The Oxford Handbook of Cognitive Linguistics. Oxford: Oxford University Press, 697-725.

Kövecses, Z. (2005). Metaphor in Culture. Universality and Variation. Cambridge: Cambridge University Press.

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

Schneider, E. (2007), Postcolonial English. Cambridge: Cambridge University Press.

Sharifian, F. (2015). Cultural Linguistics and world Englishes. World Englishes, 34 (4), 515-532.

Skandera, P. (2003). Drawing a Map of Africa: Idiom in Kenyan English. Tübingen: Narr.

Wolf, H.-G. & Polzenhagen, F. (2007a). Fixed expressions as manifestations of cultural conceptualizations: Examples from African varieties of English. In P. Skandera (Ed.), Phraseology and Culture in English. Berlin: Mouton de Gruyter, 399-435.

Wolf, H.-G. & Polzenhagen, F. (2007b). Culture-specific conceptualisations of corruption in African English: Linguistic analyses and pragmatic applications. In

F. Sharifian & G. B. Palmer (Eds.), Applied Cultural Linguistics. Amsterdam: Benjamins, 125-168.
Wolf, H.-G. & Polzenhagen, F. (2009). World Englishes. A Cognitive Sociolinguistic Approach. Berlin: Mouton de Gruyter.

• • •

Anna Cichosz
University of Łódź
an.cichosz@gmail.com

Non-inversion after þa in Old English prose: a corpus study

The Old English adverb *þa*, when placed in the clause-initial position, as in (1), causes regular inversion of all subject types, including personal pronouns, which in other contexts generally stay non-inverted.

(1) *þa* *ge-mette* **he** *sceaðan*
then met he robbers
'… then he met robbers…' (ÆLS 31.151) (after Pintzuk 1999: 91)

Thus, *þa* belongs to a very limited set of elements, referred to as 'operators' in generative studies of OE syntax (Ringe & Taylor 2014: 400), which cause inversion of pronominal subjects (Fischer et al. 2000; Haeberli 2002; Pintzuk 1999; Kroch & Taylor 1997). Nonetheless, non-inversion after *þa* is reported to take place in some specific syntactic contexts. As noticed by Mitchell, "[c]lauses in which *þa* or *þonne* follows conjunctions like *ac*, *forðæm,* and *ond*, or interjections like *efne* and *hwæt*, must be considered separately, because of the possible influence of these words on the element order" (Mitchell 1985: §2547, fn. 95). What is more, non-inversion after *þa* is said to be common when *þa* is combined with some other clause-initial element (Koopman 1998; Allen 1995: 36), e.g. *þa sona* ('immediately then').

More recent corpus-based studies confirm that the VS pattern is in general relatively infrequent in conjunct clauses (i.e. main clauses introduced by the coordinating conjunctions *ond* and *ac*) (Bech 2016) and in main clauses preceded by the interjection *hwæt* (Walkden 2013; Cichosz (forthcoming)), as in (2) and (3) where *þa* is immediately followed by the subject.

(2) *&* *ða* **Drihten** *eowre spræca gehyrde*
and then Lord your speech heard
'And then the Lord heard your speech' (cootest,Deut:1.34.4495)

(3) *Hwæt ða* **Drihten** *arærde micelne wind*
what then Lord raised up great wind
'Behold, then the Lord raised up a great wind' (cocathom2,ÆCHom_II._37: 278.205.6272)

However, alternative patterns as in (4) and (5) also exist and so far there has been no comprehensive corpus-based investigation of the contexts in which there is variation between the patterns *þa*-SV and *þa*-VS in OE prose.

(4) *Hwæt ða asprang* **micel oga** *on eallum þam folce*
what then arose great fear on all the people
'What then there was great fear among all the people' (cocathom1, ÆCHom_I,_33:459.13.6562)

(5) *and þa comon* **his leorningcnihtas**
and then came his disciples
'And then his disciples came' (coaelhom, ÆHom_5:205.806)

This study, based on the York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE) (Taylor et al. 2003), searched by means of the CorpusSearch 2 application (Randall et al. 2005-2013), aims to identify all the contexts in which *þa* does not cause regular SV inversion, check the frequency of the alternative patterns as in (4) and (5) (which, as preliminary results show, are very well-attested), and uncover the factors underlying this variation. The ultimate goal of the study is to deepen our knowledge of the OE V-2 phenomenon by establishing the reason(s) why *þa* behaves in a different way on its own and when accompanied by other, mostly extra-clausal elements such as coordinating conjunctions and interjections, which by definition should play "no part in the syntax of the sentence" (Mitchell 1985: §1234).

References

Allen, C. L. (1995). Case Marking and Reanalysis. Oxford: Oxford University Press.

Bech, K. (2016). Old truths, new corpora: revisiting the word order of conjunct clauses in Old English. English Language and Linguistics, 1-25.

Cichosz, A. (forthcoming). The constituent order of hwæt-clauses in Old English prose. Journal of Germanic Linguistics.

Fischer, O., van Kemenade, A., Koopman, W. & van der Wurff, W. (2000). The Syntax of Early English. Cambridge: Cambridge University Press.

Haeberli, E. (2002). Observations on the loss of Verb Second in the history of English. In C. J-W., Zwart & W. Abraham (Eds.), Studies in comparative Germanic syntax. Amsterdam / Philadelphia: John Benjamins, 245-272.

Koopman, W. (1998). Inversion after single and multiple topic in Old English. In J. Fisiak & Marcin Krygier (Eds.), Advances in English historical linguistics. Berlin: Mouton de Gruyter, 135-49.

Kroch, A. & Taylor, A. (1997). Verb Movement in Old and Middle English: Dialect Variation and Language Contact. In A. van Kemenade & N. Vincent (Eds.),

Parameters of Morphosyntactic Change. Cambridge: Cambridge University Press, 297-325.
Mitchell, Bruce. (1985). Old English Syntax. Oxford: Clarendon Press.
Pintzuk, S. (1999). Phrase Structures in Competition: Variation and Change in Old English word order. New York: Garland.
Randall, B., Kroch, A. & Taylor, A. (2005–2013). CorpusSearch 2. (Available online at http://corpussearch.sourceforge.net/CS.html).
Ringe, D. & Taylor, A. (2014). The Development of Old English. Oxford: Oxford University Press.
Taylor, A., Warner, A., Pintzuk, S. & Beths, F. (2003). The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE). Department of Linguistics, University of York. Oxford Text Archive. (Available online at http://www-users.york.ac.uk/~lang22/YcoeHome1. htm)
Walkden, G. (2013). The status of hwæt in Old English. English Language and Linguistics 17, 465-488.

• • •

Claudia Claridge
University of Augsburg
claudia.claridge@philhist.uni-augsburg.de

Ewa Jonsson
Uppsala University and Mid-Sweden University
Ewa.Jonsson@engelska.uu.se

Merja Kytö
Uppsala University
Merja.Kyto@engelska.uu.se

A little something goes a long way: The downtoner (*a*) *little* in the Old Bailey Corpus

Various intensifier studies (e.g. Ito & Tagliamonte 2003, Tagliamonte 2008) have noted the dominance of very few forms. If one checks downtoners in the Old Bailey Corpus (OBC 2.0), comprising ca. 24 million words spoken in a courtroom context in the period 1720-1913, it turns out that *little* is by far the most frequent downtoner (with the exception of the multifunctional intensifier *quite*) with around 8,000 occurrences. Therefore this contribution will be entirely devoted to the structural and functional profile of (*a*) *little* in Late Modern English speech-related data (we will leave the other

downtoners for a later paper); this period and our source, OBC, have so far been largely neglected in intensifier and especially in downtoner studies.

The two downtoners *little* and *a little* can function as minimizer and diminisher respectively, and also in negative litotic contexts (Quirk et al. 1985: 598, Bolinger 1972: 131, 234). Stoffel (1901: 131) further mentions the variant *a leetle* with emphasized long vowel to express "the very smallest degree". Partly depending on the forms (+/- article), they can modify nouns, adjectives, and verbs, but with certain restrictions, such as *little* mostly with comparatives/past participles and mental verbs (Bolinger 1972: 50f). (*A*) *little* may have quantity/frequency/duration and diminutive meanings, which need to be distinguished from the degree meaning most relevant here; this partly goes together with different syntactic uses and positions (e.g. emphatic front position and inversion). Modern *little* seems to be more open than other types to being itself intensified. We therefore seek to answer the following questions:

- What are the targets that speakers in the courtroom modify by using (*a*) *little* (nouns, verbs, adjectives, potentially even adverbs)? Are the restrictions noted for modern usage already in evidence or emerging?
- How do the modification patterns correlate with the different meanings and (pragmatic) functions? In which syntactic contexts are degree meanings most prominent?
- What are the distributions of the degree forms across various types of speakers with regard to speakers' social (e.g. gender and rank) and functional (e.g. judge, witness) roles? Which are the most innovative/conservative types of users in sociolinguistic respects?

Comparisons will also be drawn to the results of our previous work on *a bit* (Claridge & Kytö 2014), whose uses partly overlap with *a little* but which is a younger form. It may be assumed that (*a*) *little* is, in comparison, more established in the degree function.

References

Bolinger, D. (1972). Degree Words. The Hague/Paris: Mouton.

Claridge, C. & Kytö, M. (2014). "'You are a bit of a sneak': Exploring a degree modifier in the Old Bailey Corpus". In Hundt, Marianne (ed.), Late Modern English Syntax. Cambridge: Cambridge University Press, 239-268.

Ito, R. & Tagliamonte, S. A. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. Language in Society 32: 257-279.

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

Stoffel, C. (1901). Intensives and Down-toners. A Study in English Adverbs. Heidelberg: Winter.

Tagliamonte, S. A. (2008). So different and pretty cool! Recycling intensifiers in Toronto, Canada. English Language and Linguistics 12 (2): 361-394.

• • •

Peter Crosthwaite
University of Hong Kong
drprc80@hku.hk

Successful and unsuccessful language features of L2 group oral academic tutorial discussion: A learner corpus approach to language assessment.

In ESL contexts where English is the medium of tertiary instruction (as in Hong Kong), mastery of a range of rhetorical and interactional strategies in L2 English is required for successful English for Academic Purposes (EAP) discourse, with competence required of a variety of register-appropriate meta-linguistic stance devices used to 'stamp their personal authority or beliefs onto their arguments' (Hyland 2016:247), alongside clear, fluent L2 production. However, due to the time spent preparing for examinations in a 'competitive exam-oriented system' (Kennedy 2002:439), freshman undergraduates have had little opportunity to develop competence in oral academic engagement (Hyland 2016).

In response, the use of peer-to-peer / group oral L2 EAP assessments over the use of one-on-one oral interviews is now increasingly common. Yet qualitative studies (PRESENTER, submitted) have shown that raters vary widely in terms of their perception of successful group academic oral performance when grading. Learner corpora are now increasingly used for the purposes of language assessment (Taylor & Barker 2008; Callies & Götz 2015), yet few corpus-based studies have characterised how group oral EAP production is positively (or negatively) assessed by teacher raters. Such an enquiry would help unlock the ‚hidden curriculum' that exists between which linguistic features are taught on an EAP course and the actual features that raters use to make grading decisions.

The overriding research question is:

Which linguistic features contribute the most to the perceived success of L2 learner performance in an L2 academic group oral assessment context?

In this paper, a corpus of 59, 20-25-minute long, 5-person group oral EAP assessments spanning 20 hours and 150,309 words of L1 Hong Kong L2 English learner was constructed. The data, graded by teacher-raters as between A to C grades using an in-house can-do scale for successful academic stance, interaction and comprehensibility, was annotated for grade, 22 error types, an exhaustive range of interactive and interpersonal metadiscourse (following Hyland 2005) and a range of temporal, prosodic, lexical and syntactic markers (or ‚fluencemes') of (dis)fluency (following Götz 2013). The results suggest that successful, frequent use of metadiscourse (particularly engagement markers and topic shifts) is the primary indicator of raters' positive evaluation of student performance in L2 academic tutorial discussion alongside temporal fluencemes (namely speech rate per minute). Identical repeats and errors of idiom / collocation are salient to raters' negative appraisals, while other L2 errors and other individual prosodic (e.g. pauses), lexical (e.g. reformulations) and syntactic flu-

encemes (e.g. interrupted structures, dependent clauses) are less important to raters' positive (or negative) evaluations.

The detailed cross-sectional data afforded via this kind of corpus analysis serves as quantitative evidence of the linguistic features involved in grading decisions across the rubric, which can (and should) be used in discussions of standardisation and moderation for the raters involved.

References

Callies, M., & Götz, S. (Eds.). (2015). Learner Corpora in Language Testing and Assessment (Studies in Corpus Linguistics, Vol. 70). Amsterdam: John Benjamins Publishing Company.

Götz, S. (2013). Fluency in Native and Nonnative English Speech (Studies in Corpus Linguistics, Vol. 53). John Benjamins Publishing.

Hyland, K. (2005). Metadiscourse. New York: John Wiley & Sons, Inc.

Hyland, K. (2016). Writing with attitude: Conveying a stance in academic texts. In E. Hinkel (Ed.), Teaching English Grammar to Speakers of Other Languages (pp.246-265). London: Routledge.

Kennedy, P. (2002). Learning cultures and learning styles: Myth-understandings about adult (Hong Kong) Chinese learners. International Journal of Lifelong Education, 21(5), 430-445.

Taylor, L., & Barker, F. (2008). Using corpora for language assessment. In E. Shohami & N.H. Hornberger (eds.) Encyclopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment (pp. 241-254). New York: Springer.

• • •

Sandra C. Deshors
Michigan State University
sandracdeshors@gmail.com

Paula Rautionaho
University of Tampere
paula.rautionaho@uta.fi

The effect of grammatical contexts on the progressive vs. non-progressive alternation across World Englishes

This paper offers a multifactorial corpus-based analysis of progressive marking that contrasts native English (ENL) to two Asian Englishes (ESL) and Dutch English (EFL). Specifically, we model the constructional choices (progressive vs. non-progressive) across Englishes based on several linguistic predictors simultaneously and assess how

speakers' linguistic choices are influenced by the combined effects of their native language and different writing contexts (i.e. genre). Although a number of linguistic factors are known to influence progressive marking, previous research investigated these factors in relation to the progressive by itself rather than (i) comparing the impact those factors may have on the progressive vs. non-progressive alternation and (ii) assessing how several factors determine, simultaneously, speakers' constructional choices. Further, the lack of EFL multi-genre corpora so far prevented the exploration of genre effects in progressive usage. However, the recently compiled multi-genre Corpus of Dutch English (NL-CE) goes some way to filling this gap. Based on 6,183 progressive and non-progressive verbs constructions from five comparable corpora (the Great Britain, USA, India and Singapore sections of the International Corpus of English, and the NL-CE), across seven genres and annotated for tense, modality, semantic domain, voice, genre and variety, we ran a logistic regression analysis to determine which factors cause different English speaking populations to differ in their constructional choices and in which specific writing contexts. The model correctly predicts speakers' choices relatively strongly (C=0.82) and all predictors were significant. Overall, semantic domains emerge as contextual features that influence writers' constructional choices regardless of their English variety and their written genre. Specifically, Existence verbs, Aspect-Causative verbs and verbs denoting a mental process most significantly influence writers' choices. Because those domains transcend varieties and genres, they emerge as core determining factors in writers' constructional choices. In addition, genre effects are stronger than variety effects: while the latter are limited to the TENSE.MODALITY factor, with GENRE, significant variation occurs with AKTIONSART, TENSE.MODALITY and VOICE. Interestingly, although both VARIETY and GENRE lead to significant variation, their combined effects does not yield any deviant usage pattern; so those two different types of effects should not be assimilated. In the context of the ENL-ESL-EFL continuum, these results support the recent trend to explore the fuzzy boundaries of ESL and EFL in progressive marking (Hundt & Vogel 2011, Meriläinen et al. in print) in that American and Dutch Englishes yield similarities while ESL and Dutch English remain distinct. Finally, while Meriläinen et al. (in print) propose substrate transfer as the most likely explanation for the variation found between ENLs and ESLs, such variation is not found when the focus shifts to grammatical conditioning. Therefore, substrate transfer may not be a suitable explanation for the constructional choices witnessed in this study; rather, our results point towards universality of the progressive vs. non-progressive alternation in the ENL-ESL-EFL continuum. In that respect, our results support Sharma's (2009) view that a close scrutiny of the semantic and grammatical conditioning of the progressive is necessary in order to eliminate substrate/transfer effects and to ascertain its universality.

References

Hundt, M. & Vogel, K. (2011). Overuse of the progressive in ESL and learner Englishes - fact or fiction? In J. Mukherjee & M. Hundt (Eds.), Exploring Second-Language

Varieties of English and Learner Englishes: Bridging a Paradigm Gap. Amsterdam: John Benjamins Publishing Company, 145-165.
Meriläinen, L., Paulasto, H. & Rautionaho, P.. (In press). Extended use of the progressive in Inner Circle, Outer Circle and Extending Circle Englishes. To appear in M. Filppula, J. Klemola, A. Mauranen & S. Vetchinnikova (Eds.), Changing English: Global and Local Perspectives. Berlin: De Gruyter Mouton.
Sharma, D. (2009). Typological diversity in New Englishes. English World-Wide 30 (2), 170-195.

• • •

Florian Dolberg
Johannes Gutenberg-Universität Mainz
dolberg@uni-mainz.de

Antonina Goltsche
Johannes Gutenberg-Universität Mainz
agoltsch@students.uni-mainz.de

***Go* (*and*) *look at* register distribution on both sides of the pond: usage difference between *go*-V and *go-and*-V in British vs American English**

Go-and-V and *go-V* is a pair of rather similar double-verb constructions[1][1]:

1. *Go and get* a plate for your dad. (BNC, Spoken, KB9, S_conv)
2. Let's *go see* a movie. (COCA, FIC, FantasySciFi)

There is some debate about whether the latter developed from the former (cf. e.g. Zwicky 1969; Shopen 1971; Carden & Pesetzky 1977; Wulff 2006; Nicolle 2007, 2009), or whether *go-V* rather originated from two consecutive imperatives (cf. e.g. Visser 1969; Pullum 1990:236; Bachmann 2015). The interpretation that *go-V* and *go-and-V* are synonymous has been rejected, however how they differ is not established. For instance, Wulff found that *go-and-V* and *go-V* differ semantically: the former "gains an event-like interpretation", while the latter "is inherently atelic" (2006: 121). According to Nicolle (2007, 2009), though, *go-V* grammaticalised from *go-and-V* and now differs from it by signalling subjective construal (cf. Langacker 1999:149), following the "tendency in many grammaticalized constructions towards increased subjectivity"

1 'Construction' is employed following Goldberg (2006: 5). V signifies a slot in the construction into which a verb is inserted.

(Nicolle 2009: 204). Bachmann (2015) provides evidence from novels and short stories in the COHA suggesting *go-and-V* to be chiefly used in infinitival contexts throughout the last two centuries, while *go-V* expanded from imperative constructions to infinitival ones. Finally, Flach (2015, 2017) provides corpus analyses demonstrating that both constructions in both varieties essentially always serve the same set of broadly hortatory functions (e.g. suggestions, requests, commands, encouragements, intentions, etc.), the difference being merely "a less salient hortatory element" (Flach 2015: 248) in *go-and-V* as compared to *go-V*. These diverging interpretations call for a register analysis to shed more light on the issue.

Turning to intervarietal distribution, Eastwood (2005: 147) asserts that *go-V* is the American variant of the British *go-and-V* construction, which Wulff (2006: 102) rejects on the basis of 454 *go-V* tokens she found in the BNC[2].[2] Synchronic (Mittmann 2004: 120-121) and diachronic (Bachmann 2015) corpus-analyses reveal that *go-V* is indeed preferred over *go-and-V* in American English (AE), while British English's (BE) preference is reversed. Despite these advances, the dissimilar distribution and function(s) of this pair of constructions in the two major varieties of English and their registers remains unascertained.

The present paper aims to bridge this gap with a contrastive study of BE and AE, comparing the frequencies of the two constructions in the subsections of the BYU-BNC and the COCA. It also reappraises the aforementioned conflicting interpretations (cf. Wulff 2006; Nicolle 2009; Bachmann 2015) by supplementing them with register analyses: distinctive collexeme analyses (Gries & Stefanowitsch 2004) elucidate which different collexemes are attracted to/repelled by the V-slot in the two constructions in the different registers of the two varieties, providing a more fine-grained picture regarding differences and overlaps in function(s) and usage preference.

Preliminary results corroborate previous research, showing that the two most strongly attracted collexemes are *get* and *see* in both varieties and both constructions. However, most AE instances obtain from fiction, whereas most BE instances occur in speech, suggesting that the two constructions display not only different ratios in BE and AE, but also may serve a different set of functions in the two varieties, reflected in different frequencies and distinctive collexemes across registers and varieties.

References

Bachmann, I. (2015). Has go-V ousted go-and-V? A study of the diachronic development of both constructions in American English. In: H. Hasselgård, J. Ebeling & S. Oksefjell Ebeling (eds.), Corpus Perspectives on Patterns of Lexis. Amsterdam: John Benjamins, 91-111.

Carden, G. & Pesetsky, D. (1977). Double-verb constructions, markedness, and a fake co-ordination. In Papers from the 13th Regional Meeting of the Chicago Linguistic Society. Chicago: Chicago Linguistic Society, 82–92.

Davies, M. (2004-). British National Corpus: 100 million words, 1980s-1993. http://corpus.byu.edu/bnc/

2 Wulff (2006) does not specify which BNC-version she used.

Davies, M. (2008-). The Corpus of Contemporary American English: 520 million words, 1990-present. http://corpus.byu.edu/coca/
Eastwood, J. (2005). Oxford Guide to English Grammar. 9th Ed. Oxford: Oxford University Press.
Flach, S. (2015). Let's go look at usage: A constructional approach to formal constraints on go-VERB. In T. Herbst & P. Uhrig (eds.), Yearbook of the German Cognitive Linguistics Association 3(1), 231–252.
Flach, S. (2017). '*She goes listens to a talk on serial verbs' – Was uns Konstruktionsnetzwerke über formale Beschränkungen sagen können. Paper presented at JGU Mainz, January 23rd 2017.
Gries, S. T. & Stefanowitsch, A. (2004). Extending collostructional analysis A corpus-based perspective on 'alternations'. International Journal of Corpus Linguistics 9(1), 97-129.
Goldberg, A. E. (2006). Constructions at Work: The Nature of Generalization in Language. Oxford: Oxford University Press.
Langacker, R. W. (1999). Losing control: Grammaticization, subjectification, and transparency. In A. Blank & P. Koch (eds.), Historical Semantics and Cognition [Cognitive Linguistics Research 13]. Berlin: Mouton de Gruyter, 147–175.
Mittmann, B. (2004). Mehrwort-Cluster in der englischen Alltagskonversation. Tübingen: Gunter Narr.
Nicolle, S. (2007). The grammaticalization of tense markers: A pragmatic reanalysis. In: L. de Saussure, J. Moeschler & G. Puskas (eds.), Tense, Mood and Aspect: Theoretical and Descriptive Issues [Cahiers Chronos 17]. Amsterdam/New York: Rodopi, 47–65.
Nicolle, S. (2009). Go-and-V, come-and-V, go-V and come-V: A corpus-based account of deictic movement verb constructions. English Text Construction 2(2), 185–208.
Pullum, G. (1990). Constraints on intransitive quasi-serial verb constructions in modern colloquial English. Ohio State Working Papers in Linguistics 39:218-239.
Shopen, T. (1971) Caught in the act. In Seventh Regional Meeting, Chicago Linguistic Society. April 16-18: 254-263.
Stefanowitsch, A. & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. International Journal of Linguistics 8: 209-243.
Visser, F. Th. (1969). An Historical Syntax of the English Language. Part three, first half: Syntactical units with two verbs. Leiden, E. J. Brill.
Wulff, S. (2006). Go-V vs. go-and-V in English: A case of constructional synonymy? In S. Th. Gries & A. Stefanowitsch (eds), Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis. Berlin: Mouton de Gruyter: 101-125.
Zwicky, A. M. (1969). Phonological constraints in syntactic descriptions. Papers in Linguistics 1, 411–463.

• • •

Matthias Eitelmann
Johannes Gutenberg-University Mainz
eitelman@uni-mainz.de

Dagmar Haumann
University of Bergen
Dagmar.Haumann@uib.no

Variation, Loss and Renewal: P-Elements as Transitivizers and Transitivity Enhancers

While morphophonological marking of transitivity was lost in the transition from Old English to Middle English (cf. García García 2012), leading to the so-called "floating transitivity of English verbs" (Jespersen 1927: 319) and a high amount of labile verbs, later stages of English held on to morphosyntactic means to augment a verb's argument structure, thereby enhancing its transitivity. Three strategies have emerged in this context, which have been affected by variation, loss and renewal to differing degrees, and all of which concern transitivizing elements such as (traditional) prefixes, particles/prepositions, henceforth P-elements (cf. e.g. Jackendoff 1973, den Dikken 1995, Svenonius 2003). The first of these strategies, which involves only affixal P-elements, such as the transitivizing prefix *be-* (1a), vanished eventually, with a last heyday of *be*-derivatives in Early Modern English (1b & 1c):

(1) a. Þu ellþeodig usic woldest on þisse folcsceare facne ***be***syrwan, synnum ***be*** smitan [DOE]
you foreign us wanted in this nation treachery.INST be-deceive, sins.INST be-smear
b. 'Twere a good deed, to..***be***snowball him with rotten egges. [OED]
c. Her old Lover … was ready to ***be***pisse himselfe for feare. [EEPF]

Conversely, the second employs both the prefix and the post-verbal particle, such as Old English and Middle English *of-* in (2) and (3):

(2) a. & slog aldormonnes esne & ***of***cearf his earlipprica. [DOE]
& slew nobleman.GEN slave & off-carved his earlobes
b. & cearf ***of*** heora handa & heora nosa. [DOE]
& carved off their hands & their noses
(3) a. Lady, thy sleve thou shalte ***of***-shere [MED]
b. so huge A stroke geuyng hym was tho That quite clene the arme share ***off*** throughtly. [MED]

A third strategy emerged in analogy to the aforementioned post-verbal particle variant, exclusively coming as such and lacking a prefixed counterpart in Present-day

English, e.g. *off* and *away*, which transitivize canonically intransitive verbs such as *sleep* and *whisper* (4a,b):

(4) a. The Sleeping Beauty diet, where you're heavily sedated and you sleep ***off*** the pounds for several days. [COCA]
b. Some of her happiest memories were of working side by side with Michael O'Toole and listening to him whistle the chore ***away***. [COCA]

Against this backdrop, our paper empirically investigates the diachronic trajectories of morphosyntactically complex transitive verbs consisting of a verbal element and a transitivizing P-element. Our corpus-based study, with a focus on *be-*, *off* and *away*, is of interest for three reasons: first, the empirical investigation of the transitivizing strategies (1) – (4) reveals diachronic persistence of the morphosyntactic means employed to induce/enhance transitivity. Second, the diachrony of P-elements as transitivizers is reminiscent of cyclic change, with the particle construction progressively increasing its productivity (see e.g. Cappelle 2007) and the prefix variant recently staging a humble comeback (see also Diemer 2014). Last, but not least, the prosodic, lexical and/or syntactic constraints that, diachronically, regulate the realization of the P-element as prefix or particle, and if the latter, in verb-(non)adjacent position (cf. e.g., den Dikken 1995, Elenbaas 2007, Los et al. 2012, Thim 2012)

Corpora

[CME/MED] *Corpus of Middle English Prose and Verse/Dictionary of Middle English. 2006.* Available online at https://quod.lib.umich.edu/c/cme

[COCA] Davies, M. (2008-). *The Corpus of Contemporary American English: 450 million words, 1990–present.* Available online at http://corpus.byu.edu/coca/.

[DOE] *Dictionary of Old English Corpus.* (2000). Centre for Medieval Studies, University of Toronto.

[EEPF] *Early English Prose Fiction.* (1997). Chadwyck-Healey, Cambridge.

[OED] *Oxford English Dictionary.* (2016). Oxford University Press. Available online at http://www.oed.com/

References

Cappelle, B. (2007). When 'Wee Wretched Words' Wield Weight: The impact of verbal particles on transitivity. In M. Nenonen & S. Niemi (Eds.), Collocations and Idioms 1: Papers from the First Nordic Conference on Syntactic Freezes, Joensuu, Finland, 19-20 May 2006. Joensuu: University of Joensuu, 41-54.

Dikken, M. den (1995). Particles: On the Syntax of Verb-Particle, Triadic, and Causative Constructions. Oxford University Press.

Diemer, S. (2014). What happened to the English prefix, and could it stage a comeback? In K. Davidse, C. Gentens, L. Ghesquière & L. Vandelanotte (Eds.), Corpus Interrogation and Grammatical Patterns. Amsterdam: Benjamins, 35-55.

Elenbaas, M. (2007). The Synchronic and Diachronic Syntax of the English Verb-Particle Combination. Ph.D. thesis, Radboud Universiteit Nijmegen.

Garciá Garciá, L. (2012). Morphological causatives in Old English: the quest for a vanishing formation. Transactions of the Philological Society, 110, 122-148.
Jackendoff, R. (1973). The base rules for prepositional phrases. In S. R. Anderson & P. Kiparsky (Eds.), A Festschrift for Morris Halle. New York, NY: Holt, Rinehart, and Winston, 344-356.
Jespersen, O. (1927). A Modern English Grammar on Historical Principles. Vol. 3: Syntax, Part 2. London: George Allen & Unwin Ltd.
Los, B., C. Blom, G. Booij, M. Elenbaas, & A. Van Kemenade (2012). Morphosyntactic Change: A Comparative Study of Particles and Prefixes. Cambridge: Cambridge University Press.
Svenonius, P. (2003). Limits on P: filling in holes vs. falling in holes. Nordlyd: Proceedings of the 19th Scandinavian Conference of Linguistics 31, 431-445.

• • •

John Flowerdew
Lancaster University, University of London
johnflowerdew888@gmail.com

Meilin Chen
City University of Hong Kong
meilinchen8388@gmail.com

The use of general academic and discipline-specific corpora for research writing: introducing data-driven learning to PhD students in Hong Kong

There is no doubt that the advent of electronic corpora has revolutionised many areas of linguistic research, including applied disciplines such as language teaching, reference publishing, machine translation, and speech recognition. In language teaching, the use of corpora is no longer restricted to researchers and teachers; in the past 30 years or so there has been a growing research trend of investigating the direct use of corpora by language learners, referred to as data-driven learning (DDL, [Johns, 1991]). Indeed, Boulton and Cobb's (2017) comprehensive review identified more than 100 studies reporting on DDL endeavours to date. Notable DDL courses reported on in the field of academic writing include Lee and Swales's (2006) DDL course for a small group of doctoral students at the University of Michigan; the corpus-assisted academic writing course run by Maggie Charles (2007, 2010, 2012, 2015) at Oxford University; similar courses run by Viviana Cortes (2014) for several years at Georgia University; and a course run by Lynne Flowerdew (2015) at the Hong Kong University of Science and Technology. Most of these interventions, however, were small-scale operations,

often in experimental conditions (Boulton & Wilhelm 2006; Leńko-Szymańska & Boulton 2015) and the DDL approach has not been widely disseminated in mainstream language pedagogy. As Leech (1997: 2) pointed out nearly 20 years ago, the "trickle down" effect from research to teaching may be slower than expected

This paper reports on a territory-wide project in Hong Kong that aimed to disseminate the DDL approach among postgraduate research students (MPhil and PhD) to facilitate research writing. A half-day workshop was delivered more than 20 times at six of the eight government-funded Hong Kong universities. In total, nearly 500 students attended the workshop, accounting for 6.7% of the research degree community in Hong Kong (n =7,097, based on the 2015/16 enrolment information provided on the official website of the University Grants Committee, an advisory body for the government responsible for determining the funding of academic programmes in the eight universities). The workshop, which consisted of three parts, introduced different types of corpora to the students. In Part 1, students learned how to use the academic component of the BNCweb corpus to polish their writing at both lexico-grammatical and discourse levels. During Part 2, they were introduced to a discipline- and section-specific corpus (including seven sub-corpora) of research articles with AntConc (Laurence 2016). Adopting the genre- and corpus-based dual approach (Charles, 2007; L. Flowerdew 2005, 2009, 2015), students analysed the discourse moves of certain sections of research articles and their linguistic realisations. In Part 3, the final part, students went through the process of creating their own personal corpora under the teacher's guidance. Selected activities from the workshop will be demonstrated during the presentation.

Results from the post-workshop survey show that the great majority of the students were not familiar with corpora or DDL before attending the workshops, and they greatly appreciated the value of this new approach. Their evaluation of the workshops will be presented and discussed. The findings from this project may shed further lights on the value and feasibility of spreading the DDL approach in institutions at research degree level.

References

Anthony, L. (2016). AntConc 3.4.4. [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Boulton. A. & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. Language Learning, 67(2).

Boulton, A. & Wilhelm, S. (2006). Habeant Corpus—they should have the body. Tools learners have the right to use. ASp. la revue du GERAS, (49-50), 155-170.

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

Charles, M. (2011). Using hands-on concordancing to teach rhetorical functions: Evaluation and implications for EAP writing classes. In A. Frankenberg-Garcia, L. Flowerdew & G. Aston (Eds.), New Trends in Corpora and Language Learning (pp. 81-104). London/New York: Continuum International Publishing Group.

Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus building. English for Specific Purposes, 31(2), 93–102.
Charles, M. (2015). Same task, different corpus: The role of personal corpora in EAP classes. In A. Leńko-Szymańska & A. Boulton (Eds.), Multiple Affordances of Language Corpora for Data-driven Learning (pp. 129-154). Amsterdam: John Benjamins.
Cortes, V. (2014). Genre analysis in the academic writing class: with or without corpora? Quaderns de Filologia-Estudis Lingüístics, 16, 65-80.
Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. English for Specific Purposes, 24 (3), 321–332.
Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. International Journal of Corpus Linguistics, 14(3), 393-417.
Flowerdew, L. (2015). Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. Journal of English for Academic Purposes, 20, 58-68.
Johns, T. (1990). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. CALL Austria, 10, 14-34.
Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), Teaching and Language Corpora (pp. 11–23). Harlow: Addison Wesley Longman.
Leńko-Szymańska, A. & Boulton, A. (2015). Introduction: Data-driven learning in language pedagogy. In A. Leńko-Szymańska & A. Boulton (Eds.), Multiple Affordances of Language Corpora for Data-driven Learning (pp. 1-14). Amsterdam: John Benjamins.

• • •

Dana Gablasova
Lancaster University
d.gablasova@lancaster.ac.uk

A corpus-based approach to the expression of subjectivity in L2 spoken English: The case of '*I* + verb' construction

Subjectivity in language refers to how speakers express "their perceptions, feelings and opinions in discourse" (Scheibmann 2002) and to the linguistic features and structures that enable "self-expression in the use of language" (Lyons 1994; Benveniste, 1958). The expression of subjectivity is an important component of pragmatic ability as it is closely related to how speakers communicate politeness (e.g. boosting or downplaying one's involvement) or their stance in interaction (Reilly et al. 2005).

In order to contribute to our understanding of how spoken, interactive production develops in learner language, the paper investigates the expression of speaker's subjective involvement in the 'I + verb' construction in spoken production of L2 English by speakers of different levels of English proficiency. In particular, it focuses on two prominent lexical categories of verbs (Biber et al. 1999; Scheibman 2002; Levin 1993) that occur in this construction – emotive (e.g. *love, need, wish*) and cognitive/epistemic (e.g. *believe, think, suppose*) verbs.

The study uses the Trinity Lancaster Corpus (TLC) of spoken L2 production (Gablasova et al. 2015) based on examinations of spoken English conducted by Trinity College London (a major examination board). L1 Spanish and Italian speakers aged over 20 years (to control for the effect of cognitive maturity on expressions of subjectivity) were selected from the TLC to represent three proficiency levels of the Common European Framework of Reference: B1 (183 speakers), B2 (170 speakers) and C1/C2 (102 speakers). All speakers participated in two interactive speaking tasks (conversations) which together lasted approximately 10 minutes. Using MonoConc Pro (Barlow 2004), all '*I* + verb' constructions in speakers' production were identified and the verbs in these constructions were categorised as emotive, cognitive or other (e.g. material, auxiliary) verbs (e.g. Biber et al. 1999; Scheibman 2002).The ANOVA was used to compare the frequency of each verb category across the three proficiency bands.

The findings show a very clear and statistically significant trend in the use of the '*I*+ verb' construction. With the increase in proficiency the frequency of emotive verbs decreased while the frequency of the epistemic verbs increased considerably. The study also identified the most frequent cognitive and emotive verbs and the trends in their use according to the proficiency level of L2 users. The study contributes to a larger discussion of the effect of lexico-grammatical competence on the development of pragmatic competence (e.g. Schauer 2013; Kasper & Rose 2002) and discusses the findings from the perspective of second language pragmatic ability.

• • •

Evelyn Gandón-Chapela
University of Cantabria/University of Vigo
evelyn.gandon@gmail.com

How long is it from the antecedent to the ellipsis site? Lexical/syntactic distance in English ellipsis

This study, as part of a completed larger project on ellipsis, undertakes a corpus-based analysis of lexical (in number of words) and syntactic distance (in number of clauses) between the antecedent clause(s) and the ellipsis site in examples of Post-Auxiliary Ellipsis (PAE henceforth) in English, using data from the Penn Parsed Corpus of Modern British English. The term PAE (Sag 1976; Warner 1993; Miller 2011; Miller and Pullum 2014) covers those cases in which a Verb Phrase (VP), Prepositional Phrase (PP), Noun Phrase (NP), Adjective Phrase (AP) or Adverbial Phrase (AdP) is omitted after one of the following licensors (those elements that permit the occurrence of ellipsis): modal auxiliaries, auxiliaries *be*, *have* and *do*, and the infinitival marker *to* (the latter believed to be a defective non-finite auxiliary verb; see Miller and Pullum 2014). This study focuses on two subtypes of PAE, namely VP ellipsis (VPE henceforth) and Pseudogapping (PG henceforth), illustrated below:

(1) I have written a squib but I think that Mary hasn't. (VPE: lexical distance: 6 words; syntactic distance: 1 clause)
(2) John is talkative but Sara is not. (VPE: lexical distance: 4; syntactic distance: 0)
(3) John kissed Sarah, and Mary did Paul. (PG: lexical distance: 4; syntactic distance: 0)

In line with the few studies which have approached this research question from an empirical perspective (Hardt 1990; Hardt and Rambow 2001; Nielsen 2005; Martin and McElree 2008), in this investigation I tackled ellipsis distance by bringing new data retrieved from a larger textually balanced electronic collection of texts. The analysis has revealed that most instances of PG occur in contexts where the antecedent is contained within the same sentence, the target of ellipsis appears in a different clause, there are no intervening clauses and, in addition, the lexical distance ranges from 0-10 words. Regarding VPE, in the vast majority of the examples the lexical distance is lower, i.e. 0-5 words. VPE differs from PG with respect to the most frequent type of boundedness established: the antecedent and the ellipsis site may either appear in different sentences or within the same sentence and in a different clause (with no intervening clauses) respectively. These data confirm that VPE has more local lexical/syntactic scope than PG since the distance metrics are shorter in the former. In addition, the results for PG corroborate Levin's (1986) findings, as PG disfavours those syntactic contexts where the pseudogapped clause is embedded. It has also been found that Hardt's (1993) hypothesis stating that lexical distance is higher in cases where the antecedent and the ellipsis site appear in different sentences is not confirmed in PG. The higher the lexical distance, the fewer instances of PG are found where the antecedent

and the ellipsis site occur in different sentences. In contrast, Hardt's (1993) hypothesis for VPE is confirmed: if lexical distance increases, the antecedent and the ellipsis site occur in different sentences in the vast majority of cases. Finally, after checking the interaction of lexical and syntactic distance, it has been found that the higher the lexical distance, the higher the syntactic distance in both PAE constructions.

References

Hardt, D. (1990). A corpus-based survey of VP ellipsis. Ms. University of Pennsylvania.

Hardt, D. (1993). Verb Phrase Ellipsis: Form, Meaning, and Processing. Ph.D. thesis, University of Pennsylvania.

Hardt, D. & Rambow, O. (2001). Generation of VP Ellipsis: A Corpus-based Approach. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Toulouse, France, 9-11 July 2001 (pp. 290–297).

Levin, N. (1986). Main Verb Ellipsis in Spoken English. Outstanding Dissertations in Linguistics. New York: Garland.

Martin, A. E. & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. Journal of Memory and Language, 58, 879–906.

Miller, P. (2011). The choice between verbal anaphors in discourse. In I. Hendrickx, S. L. Devi, A. Branco, & R. Mitkov (Eds.), Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011, volume 7099 of Lecture Notes in Artificial Intelligence (pp. 82–95). Berlin: Springer.

Miller, P. & Pullum, G. K. (2014). Exophoric VP Ellipsis. In P. Hofmeister & E. Norcliffe (Eds.), The Core and the Periphery: Data-Driven Perspectives on Syntax Inspired by Ivan A. Sag (pp. 5–32). Stanford, CA: CSLI Publications.

Nielsen, L. A. (2005). A Corpus-Based Study of Verb Phrase Ellipsis Identification and Resolution. Ph.D. thesis, King's College London.

Sag, I. (1976). Deletion and Logical Form. Ph.D. thesis, MIT.

Warner, A. (1993). English Auxiliaries: Structure and History. Cambridge: Cambridge University Press.

• • •

Gregory Garretson
Uppsala University
gregory.garretson@engelska.uu.se

Rachele De Felice
University College London
r.defelice@ucl.ac.uk

The Hillary Rodham Clinton emails: Examining world leaders through corpora

Corpora of emails open up exciting avenues of research into speech acts, specialised discourses, gendered language, and social interaction in professional environments, inter alia (De Felice 2013). However, despite email's importance and near-ubiquity over two decades, there are remarkably few publicly available email corpora. It thus behoves us to take advantage of exceptional events such as the release into the public domain of email databases, as with the Enron emails (Klimt and Yang 2004), and now Hillary Clinton's emails from her tenure as Secretary of State (US Department of State 2016). This talk presents (a) a project currently underway to turn these last into an orderly, easily searchable, publicly available corpus, and (b) the results of the first linguistic study to be carried out on this data.

Such email databases are treasure troves in several respects. Besides providing large amounts of email data (the Clinton database includes over 30,000 messages), they are rich examples of high-stakes decision-making in professional environments. In the present case, there is another reason for interest: Clinton unexpectedly lost the 2016 US presidential election, one of the few scandals to plague her campaign being her use of a private email server while Secretary of State. Despite the forced release of these emails, mistrust of Secretary Clinton's actions remained a serious problem. This project is designed to facilitate access to this historically important data by linguists and researchers in other fields.

Nevertheless, to reach that point, there are many technical challenges to overcome. The data has been released as unordered PDF files with multiple layers of metadata and featuring many redactions, making data extraction difficult. Further, the medium of email itself presents challenges such as threading (with duplicate text), boilerplate text, attachments, and multiple classes of recipients. As we work to resolve these issues, our current research focuses on a 500-email subcorpus.

The talk has two aims. The first is to describe the compilation of the corpus and discuss the more interesting aspects of the process, such as how to deal with redactions, which technical aspects of the messages to preserve, and how to determine and record the relationships between participants. The second is to present the results of a study on the 500-email subcorpus, in which we make use of the extensive background information available about the interlocutors and their relationships to each other to compare the communicative patterns in evidence across variables such as gender, hierarchical level and in-group/out-group. Specifically, we test Gilbert's (2012) claim

that at work, email is the performance of power and hierarchy, by focusing on communication (a) between members of the inner circle of the State Department and (b) between these and individuals outside the organization. The results indicate that a complex interaction of factors beyond hierarchy determines linguistic choices such as directness and informality, showing this corpus to be a unique and valuable lens through which to catch a glimpse of world leaders in action.

References

De Felice, R. (2013). A corpus-based classification of commitments in Business English. In J. Romero-Trillo (ed.), Yearbook of Corpus Linguistics and Pragmatics 2013, 153-171.

Gilbert, E. (2012). Phrases That Signal Workplace Hierarchy. In Proceedings of CSCW 2012, the 2012 ACM Conference on Computer Supported Cooperative Work.

Klimt, B. & Yang, Y.. (2004). The Enron corpus: A new dataset for email classification research. In Proceedings of ECML 2004, the 15th European Conference on Machine Learning, 217-226.

US Department of State. (2016). Freedom of Information Act Virutal Reading Room. Accessed via https://foia.state.gov/Search/Results.aspx?collection=Clinton_Email.

• • •

Sandra Götz
Justus Liebig University Giessen
Sandra.Goetz@anglistik.uni-giessen.de

Non-canonical Syntax in Varieties of English in the Indo-Pacific: A Corpus-Based Study on Fronting in South-Asian Englishes

Apart from Indian English (IndE), which constitutes the largest institutionalized second-language variety of English worldwide, English also fulfills important, yet varying, roles in India's neighboring countries Bangladesh (BgE), Nepal (NpE) and Pakistan (PkE), as well as in Sri Lanka (SLE) and the Maldives (MdE).

Previous corpus-based studies describing similarities and differences across South Asian Varieties of English (SAVEs) have mainly focused on the description of nativization processes on the lexicogrammatical level, for example in the areas of particle verbs (Schneider 2004), article use (Sand 2004) or collostructions (Mukherjee & Gries 2009; Gries & Mukherjee 2010). Previous corpus-based and quantitative studies describing syntactic aspects of SAVEs have rarely been undertaken, however, two laudable exceptions being Lange (2012), who describes in detail the (non-canonical) syntax of spoken Indian English and Winkle (2015), who compares eight spoken ENL and ESL varieties. Both studies found a particularly high frequency of fronted elements in

spoken Indian English. However, to the best of my knowledge, there has not been a study investigating systematically the parallels and disparities of syntactic patterns in different SAVEs and British English. Against this background, the present paper aims to close this research gap by taking into account fronting as a non-canonical syntactic structure in order to test 1) if fronting is a typically Indian English feature or rather a pan-South-Asian one, 2) if there are differences in form and frequency of fronted elements across different SAVEs, 3) if previous findings for spoken data (e.g. Lange 2012; Winkle 2015) also figure in written data, and 4) if the emergent structures can be explained by various linguistic and non-linguistic factors (e.g. length or syntactic complexity of the arguments, information status or "evolutionary status" (Schneider 2003) of the variety in question, etc.).

The data analysis is based on the SAVE (*South Asian Varieties of English*; cf. Bernaisch et al. 2011) corpus, a collection of texts representing acrolectal newspaper English which amounts to 6 x 3 million words (2 subcorpora at 1.5m words each) per variety. In the present paper, I would like to present the findings of study based on 1,000 sentences per variety that were manually parsed and the sentence-initial elements were annotated for several explanatory variables (e.g. LENGTH OF ELEMENT, INFORMATION STATUS (given/new), VARIETY, EVOLUTIONARY STATUS, etc.). Methodologically, I apply multifactorial regression analyses (cf. Gries 2013) in order to test for variety-specific as well as universal features of fronting in SAVEs. The results of these analyses yield some interesting findings: While it is clearly visible that a constituent is more likely to be fronted when the information is given regardless of variety (cf. also Birner & Ward 1998; Winkle 2015), there are also clear variety-specific differences between IndE, PkE, SLE and BdE on the one hand (showing a generally higher frequency of object fronting), and BrE, MdE and NpE on the other (with a higher frequency of fronted adjuncts). These findings, among others, will be discussed with regard to their implications on the norm-providing potential of SAVE as well as on the role of Indian English as a possible linguistic epicenter in South Asia.

References

Bernaisch, T., Koch, C., Schilk, M. & Mukherjee, J. (2011). Manual to the South Asian Varieties of English (SAVE) Corpus. Giessen: Justus Liebig University, Department of English.

Birner, B. J. & Ward, G.. (1998). Information Status and Noncanonical Word-Order in English. Amsterdam: John Benjamins.

Gries, S. Th. (2013). Statistics for linguistics with R. 2nd rev. and ext. ed. Berlin and New York: De Gruyter Mouton.

Gries. S. Th. & Mukherjee, J. (2010). Lexical gravity across varieties of English: an ICE-based study of n-grams in Asian Englishes. International Journal of Corpus Linguistics 15 (4), 520-548.

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

Mukherjee, J. & Gries, S. Th. (2009). Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. English World-Wide 30 (1): 27-51.

Sand, A. (2004). Shared morpho-syntactic features in contact varieties of English: article use. World Englishes 23(2), 281-298.
Schneider, E. W. (2003). The dynamics of new Englishes: From identity construction to dialect birth. Language 79(2): 233-281.
Schneider, E. W. (2004). How to trace structural nativization: particle verbs in world Englishes. World Englishes 23(2) 227-249.
Winkle, C. (2015). Non-canonical structures, they use them differently: information packaging in spoken varieties of English. PhD Dissertation. University of Freiburg.

• • •

Andrew Hardie
Lancaster University
a.hardie@lancaster.ac.uk

Plotting and comparing corpus lexical growth curves as an assessment of OCR quality in historical news data

Historical corpora which have been generated via optical character recognition (OCR) rather than reliable-but-laborious rekeying are well-known to contain "messy/noisy" data. The OCR causes a greater or lesser number of errors, whose existence poses problems for corpus analysis in terms of (a) reduced query recall and (b) a word-type's true frequency being spread across multiple (apparent, spurious) types.

A known error rate can be allowed for in interpreting results. But the extent of OCR errors in a given corpus is often not known. Historical newspaper collections, in particular, display highly heterogeneous OCR error rates, which may vary within a single document between pages/articles, between issues of one newspaper over time, and between different newspaper titles in an archive. Accurately measuring error rates can only be done by comparison to a hand-corrected gold-standard; but for sizeable collections such a gold-standard can be constructed for only small samples, and assembling a representative sample is problematised by the exact heterogeneity at issue.

Thus, OCR error rates are typically estimated by approximations, e.g. counting any token not found in a reference lexicon as an error. But this necessitates an appropriate lexicon. Moving back in history, the match between our readily accessible or creatable large lexicons and the corpora under study declines in terms of both genre and historical variety. Moreover, for many purposes we wish also to address newspaper archives in other languages than just English, amplifying this practical problem.

To characterise OCR "messiness" in historical newspaper corpora we begin with the observation that OCR errors typically generate far more *hapax legomena* than

would a non-OCR'd corpus, e.g. the London *Times* archives for 1900-1909 and 1980-1989 are both ~540MW in extent, but the former has 41 million types, the non-OCR'd latter 16 million. More elaborate quantitative models, including but not only those based upon Zipf's law (Baayen 2001:13-32), can formalise such observations regarding type/*hapax* counts. However, models positing overall parameters for a whole corpus do not permit taking account of this data's known heterogeneity.

If a graph is constructed of number of tokens observed versus count of types at intervals (say, every 10,000 tokens) a curve characteristic of *lexical growth* over the span of a given corpus emerges (see Baroni 2008:818-819). Such curves preserve corpus proximity and sequence, allowing observation by eye of progressively changing wordform profusity.

This data-visualisation technique has been applied in e.g. McEnery and Wilson's (1996/2001:173-180) study of sublanguages, whose closed lexicons create abnormally flattened curves – whereas curves for historical OCR data imply the expected abnormally open behaviour. Visual comparison of lexical growth curves among historical collections, or to modern corpora, therefore generates a good impression of the relative extent of OCR noise, and thus some estimate of how much such noise will impede analysis.

Since this method requires no reference lexicon, we can also compare OCR error incidence between collections in English and in other languages. Moreover, we can observe the effects of larger-scale heterogeneity across a corpus. This technique has been implemented within the corpus-analysis software CQPweb (Hardie 2012).

References

Baayen, R. H. (2001). Word Frequency Distributions. Dordrecht: Kluwer Academic Publishers.

Baroni, M (2008). Distributions in text. In A. Lüdeling & M. Kytö (eds), Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter, 803-822.

Hardie, A (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics 17 (3): 380-409.

McEnery, T & Wilson, A (1996, 2nd ed. 2001). Corpus Linguistics. Edinburgh University Press.

• • •

Abi Hawtin
Lancaster University
a.hawtin@lancaster.ac.uk

The British National Corpus Revisited: Developing parameters for Written BNC2014.

The Centre for Corpus Approaches to Social Science (CASS) at Lancaster University and Cambridge University Press are collaborating on the creation of a new, publicly accessible corpus of contemporary written British English: *Written BNC2014*, which together with the in-progress Spoken corpus will complete the BNC2014 initiative. This initiative aims to create a new, up-to-date corpus along the lines of the British National Corpus (BNC 1994), constructed in the early 1990s. The Written BNC1994 continues to be used even now as a proxy for contemporary British English, despite being 20 years old; a new standard, broad-coverage, widely-available corpus is needed to allow the same kind of research fostered by the Written BNC1994 to generate results that are truly representative of contemporary British English.

This presentation firstly discusses the rationale for the Written BNC2014 project. With extremely large web-crawled corpora now becoming commonplace in corpus linguistics, the question might well be asked why a new, 100 million word corpus of Written British English needs to be created. I will argue that there exist many benefits of a smaller, 'hand-made' corpus. These benefits include, but are not limited to, inclusion of data which it is not possible, or at least not straightforward, to access via a web-crawl (e.g. published books); greater certainty and control over whether the texts included in the corpus truly represent British English; and the ability to ensure that data from a wide spread of genres has been collected.

The second part of the presentation introduces the corpus construction project and provides a progress report. I will cover, first, the decisions made in the development of the corpus's composition scheme, and in particular, the divisive issue of whether the corpus design should prioritise representativeness (of contemporary language) or comparability (with the 1994 corpus). To aim for 'representativeness' would mean trying to make the corpus, as far as is possible, representative of contemporary Written British English regardless of how this would match up with Written BNC1994. Aiming for 'comparability' would mean collecting data which lines up exactly with what was included in Written BNC1994, allowing for direct diachronic comparison. We reconcile these two approaches by treating the second as a subset of the first; the corpus will aim to represent contemporary language (e.g. by including 'e-language' genres whose prevalence postdates BNC1994), but a sub-corpus will be defined which is comparable with Written BNC1994. Finally, I will present the decisions which have been made regarding the collection of the texts, along with the sampling frame designed to guide text collection for the corpus. The estimated release date for the corpus is third quarter 2018.

• • •

Benedikt Heller
KU Leuven
benedikt.heller@kuleuven.be

Tobias Bernaisch
JLU Giessen
Tobias.J.Bernaisch@anglistik.uni-giessen.de

Stefan Th. Gries
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

A corpus-linguistic account of the history of the genitive alternation in Singapore English: evidence for the transition from nativisation to endonormativity?

Singapore English is an English-as-a-second-language (ESL) variety which has progressed substantially along the evolutionary cline and – having completed the stage of nativisation around the 1970s – is currently in the phase of endonormative stabilisation. This procedural emancipation from the historical input variety British English is constituted by a complex interplay of changing attitudes towards (local) varieties of English, historically varying functional profiles of contexts of use and structural diversification of the language itself. While laudable exceptions exist, particularly this latter historical process of continuous structural diversification in outer-circle varieties has so far rather been assumed to have occurred due to present-day differences between the structures of British English and ESL varieties than empirically validated.

Thus, the focus of the present paper is the diachronic development of the genitive alternation, i.e. the *of*-genitive (as in *the food of the dog*) and the *s*-genitive (as in *the dog's food*), in Singapore English. While earlier research has produced partly diachronic accounts of genitive variability also focusing on Asian Englishes from a contemporary perspective, the diachronic development of the genitive has so far not been studied in ESL contexts. The data are several thousand data points from a stratified sample of both genitives from comparable sections of the *Corpus of Historical Singapore English* and the Singaporean component of the *International Corpus of English* (ICE) covering written Singapore English texts from several genres from the 1950s, 1960s and 1990s. The corresponding texts from the British ICE component from the 1990s serve as a present-day native-speaker reference. In accordance with earlier research, each genitive instance is annotated with phonetic (e.g. final sibilancy of possessor), semantic (e.g. animacy of possessor/possessum), syntactic (e.g. length of possessor/possessum) and pragmatic (e.g. discourse accessibility of possessor/possessum) variables.

Via *Multifactorial Prediction and Deviation Analysis with Regression*, we identify whether and how variable constraints of the genitive in Singapore English changed in the 40-year time span in comparison to present-day British English. As Singapore

English has been described to enter the phase of endonormative stabilisation in the 1970s, we expect a continuous structural emancipation to be reflected in a) a stronger structural divergence of the 1960s Singapore genitives from the 1990s genitives compared to that from the 1950s genitives, b) a historically increasing divergence between the genitive in Singapore and British English as well as c) a diachronically stable core of nouns strongly attracted by the genitive slots with an increasing number of nouns with weaker ties to the respective slots. Methodologically, this is the first MuPDAR study in which the differences between the BrE source variety and Singapore English are also statistically controlled for how BrE and Singapore English change in their own rights.

• • •

Thomas Herbst
FAU Erlangen-Nürnberg
thomas.herbst@fau.de

Corpus evidence – collostructional analysis – ColloConstruction Grammar

Traditional valency analysis (VALBU 2004, VDE 2004) and related approaches such as Pattern Grammar (Hunston & Francis 2000) provide descriptions of the complementation of verbs, adjectives, nouns, etc. in terms of formal categories or patterns. Semantically, the various slots of the construction are generally characterized in terms of abstract semantic roles at different levels of abstraction – either in terms of low-level item-specific participant roles (Herbst 2014ab) or, as in Goldberg's (1995, 2006) model of argument structure constructions, in terms of more general argument roles.

In this paper, I would like to apply the method of collostructional analysis (Stefanowitsch & Gries 2003, Gries & Stefanowitsch 2004ab) in the analysis of different slots in valency constructions of English verbs in order to explore the relations between different patterns of the same verbs. One of the issues to be discussed in this context is optionality: in traditional valency theory, a complement that can, but need not be realized is usually called an optional complement. Thus, *you* in sentence (1) below could be seen as an optional complement because the showee = person shown something does not have to be expressed, as in (2):

(1) Come on, I'll show you the way. BNC ACB 2563
(2) The wind showed no signs of abating … BNC GW3 1720

Within this line of thinking, one of the valency constructions of *show* could be described in the following way (the brackets indicating optionality):

NP shower V (NP showee) NP item shown

However, a collostructional analysis of the different uses of *show* in the BNC carried out with the help of treebank.info (Uhrig & Proisl 2012) reveals that the shower-slot and the item shown-slot have different collostructional profiles in divalent uses such as (1) and trivalent uses such as (2). It will be argued that this presents a strong argument for analysing (1) and (2) as representing two different valency constructions and an argument against the notion of optionality in the traditional sense.

Further cases to be investigated in this respect include a comparison of the collo-profiles of subjects and indirect objects in di- and trivalent uses of verbs such as *earn* as well as of the collo-profiles of active objects and passive subjects.

This paper thus aims to explore the idea of describing the slots of valency constructions in terms of so-called collo-profiles. If it can be shown that collo-profiles turn out to be of greater relevance to the description of syntactic patterns than abstractions in terms of semantic roles, then this allows far-reaching conclusions as to the mental representation of such constructions and provides evidence in favour of an exemplar-based model as proposed by Bybee (2010). In this respect, this paper is also an attempt to show how large-scale corpus analysis can provide valuable insights concerning the design of cognitive models of language.

References

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Goldberg, A. E. (1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.

Goldberg, A. E. (2006). *Constructions at Work*. Oxford & New York: Oxford University Press.

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

Gries, S. & Stefanowitsch, A. (2004b). Covarying collexemes in the into-causative. In M. Archard & S. Kemmer (eds.), *Language, Culture, and Mind*, 225–236. Stanford CA: CSLI.

Herbst, T., Heath, D., Roe, I. & Götz, D. (2004). *A Valency Dictionary of English*. Berlin & New York: Mouton de Gruyter. (= VDE)

Herbst, T. (2014a). The valency approach to argument structure constructions. In T. Herbst, H.-J. Schmid & S. Faulhaber (eds.), *Constructions – Collocations – Patterns*, 167–216. Berlin & Boston: de Gruyter Mouton.

Herbst, Th. (2014b). Idiosyncrasies and generalizations: argument structure, semantic roles and the valency realization principle. In M. Hilpert & S. Flach (eds.), *Yearbook of the German Cognitive Linguistics Association, Jahrbuch der Deutschen*

Gesellschaft für Kognitive Linguistik, Vol. II., 253–289. Berlin, München & Boston: de Gruyter Mouton.
Hunston, S. & Francis, G. (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins.
Stefanowitsch, A. & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8 (2), 209-243.
Uhrig, P., Proisl, T. (2012). Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates. *Lexicographica* 28, 141–180.
Schumacher, H., Kubczak, J., Schmidt, R. & de Ruiter, V. (2004). *VALBU – Valenzwörterbuch deutscher Verben.* Tübingen: Narr.

• • •

Mikko Höglund
Stockholm University
mikko.hoglund@english.su.se

Turo Vartiainen
University of Helsinki
turo.vartiainen@helsinki.fi

Take your hands off (of) me! **On the development of the complex preposition *off of***

The syntagm *off of* can be interpreted in at least two different ways depending on the context. The sequence can be regarded either as a single unit, the complex preposition *off of*, or as a combination of the adverbial particle *off* and the preposition *of*. In some contexts, the adverbial particle + preposition interpretation is the only possibility, as in (1), but in many, or even most, instances in which the complex preposition reading is possible, the adverbial particle + preposition analysis is also conceivable (see examples (2)–(3) below).

(1) preventing also the falling **off of** the Hair or Feathers (EEBO, 1696)

The complex preposition *off of* is semantically redundant in the sense that the simple preposition *off* can be substituted for it in most cases. This is largely due to the fact that *off* and *of* derive historically from the same lexical item (*OED*: "off" and "of"), and the 'away', or 'source', sense is present in both words. The syntagm *off of* has several similarities to other adverbial particle + preposition sequences, such as

out of (Cappelle, 2001), but it also displays many unique features that merit further investigation.

In this paper, we present a corpus-based diachronic analysis of *off of* from the 1560s to the present day. Our data show that in the 1600s *off of* started to be used in different kinds of ambiguous contexts (Diewald, 2002) that allowed for a complex prepositional reading in addition to the adverbial particle + preposition one. Based on the data from the *Early English Books Online Corpus* (EEBO), we suggest that the use of *off of* in these ambiguous contexts led to its partial grammaticalization; only partial, because the ambiguity concerning the category of *off* has never been fully resolved (see Cappelle 2001: 324). Examples (2) and (3) from the EEBO Corpus illustrate how *off of* is used in two such contexts.

(2) the Hangman will take'em **off of** our hands (EEBO, 1689)
(3) a Borough-Town 12 Miles South of Kildare […], not far **off of** the Borders of Queens-County (EEBO, 1691)

Interestingly, recent data from the *Corpus of Historical American English* show that the frequency of *off of* has more than doubled from the 1980s to the 2000s: it is as frequently used now than in its heyday in the early 18th century. Data from the *Corpus of Contemporary American English*, on the other hand, show that *off of* is particularly frequent in spoken conversation in Present-day English, a result that is consistent with spoken discourse from the 17th and 18th centuries as represented in *the Old Bailey Corpus*. These results suggest that the low frequency of *off of* in many of the historical corpora may in part be due to editorial policies and prescriptivism related to the perception of *off of* as a non-standard (Hughes and Trudgill, 1979), lower-class (Fries, 1940, p. 127) or dialectal feature (Cheshire et al. 1993: p. 77; Vasko 2010).

References

Cappelle, B. (2001). Is out of always a preposition? Journal of English Linguistics, 29 (4), 315–328.

Cheshire, J., Edwards, V. & Whittle, P. (1993). Non-standard English and dialect levelling. In J. Milroy & L. Milroy (Eds.), Real English. The grammar of English dialects in the British Isles. London: Longman, 53–96.

Davies, M. (2008–). The Corpus of Contemporary American English: 450 million words, 1990–present. http://corpus.byu.edu/coca/.

Davies, M. (2010-). The Corpus of Historical American English: 400 million words, 1810-2009. http://corpus.byu.edu/coha/.

Diewald, G. (2002). A model for relevant types of contexts in grammaticalization. In I. Wischer & G. Diewald (Eds.), New reflections on grammaticalization. Amsterdam: John Benjamins, 103–120.

Fries, C. C. (1940). American English grammar: The grammatical structure of present-day American English with especial reference to social differences or class dialects. New York: Appleton-Century-Crofts.

Huber, M., Nissel, M., Maiwald, P. & Widlitzki, B. (2012). The Old Bailey Corpus. Spoken English in the 18th and 19th centuries. www.uni-giessen.de/oldbaileycorpus/.
Hughes, A. & Trudgill, P. (1979). English accents and dialects: An introduction to social and regional varieties of British English. London: Edward Arnold.
OED Online. s.v. "of, prep." and "off, adv., prep., n., and adj.". Oxford University Press.
Vasko, A. (2010). Cambridgeshire dialect grammar. Studies in Variation, Contacts and Change in English 4. Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/04/index.html/.

• • •

It is time that this (*should*) be studied across a broader range of Englishes: a global trip around subjunctives

Marianne Hundt
University of Zürich
m.hundt@es.uzh.ch

English has a choice between modal verb *should* and a present tense subjunctive verb in subordinate *that*-clauses following mandative expressions such as *recommend, request* or *require*. Previous research has found that the subjunctive variant has seen a revival in the twentieth-century (e.g. Övergaard 1995, Leech et al. 2009 or Hundt and Gardner, 2017). This change has been led by American English (AmE) while British English (BrE) is lagging considerably behind in this revival of a conservative grammatical construction, and varieties such as Australian (AusE) and New Zealand English (NZE) are somewhat more advanced (see Hundt 1998). Studies beyond the inner circle (i.e. varieties of English as a first language) are scarce and typically look at only one second-language variety of English, comparing its text frequency with the one found in first language varieties (e.g. Indian English in Sayder 1989 or Philippine English in Schneider 2011; Peters 2009 uses evidence from Singapore and the Philippines as well as four inner-circle Englishes). On the basis of the *International Corpus of English* and supplementary evidence from NOW, the paper compares a broad range of Englishes. The evidence shows that varieties which are geographically close to or historically related to AmE show the highest proportion of subjunctives in mandative contexts, whereas other Englishes mostly (with the exception of Australian and New Zealand English) are more conservative. The historical connection with BrE is also born out by the fact that the conservative varieties also occasionally use indicatives after a mandative trigger, a variant that is not attested in AmE. The verb *be*, on the other hand, continues to be a stronghold of the subjunctive across all WEs investigated. Preliminary evidence from NOW, moreover, indicates that BrE has been catching up with AmE,

whereas the periphrastic variant with *should* finds a stronghold in some post-colonial varieties, notably Indian and Pakistani English.

References

Crawford, William J. 2009. The mandative subjunctive. In Günter Rohdenburg and Julia Schlüter, eds. One Language, Two Grammars? Differences between British and American English. Cambridge: Cambridge University Press, 257-276.Hundt, Marianne. 1998. It is important that this study (should) be based on the analysis of parallel corpora: On the use of the mandative subjunctive in four major varieties of English. In Hans Lindquist, Staffan Klintborg, Magnus Levin and Maria Estling, eds. The Major Varieties of English. Växjö: Acta Vexionensia, 159-175.

Hundt, Marianne and Anne Christine Gardner. 2017. Corpus-based approaches: Watching English change. In Laurel Brinton, ed. English Historical Linguistics: Approaches and Perspectives. Cambridge: Cambridge University Press, 96-130.

Övergaard, Gerd. 1995. The Mandative Subjunctive in American and British English in the 20th Century. Uppsala: Almquist and Wiksell.

Peters, Pam. 2009. The mandative subjunctive in spoken English. In Pam Peters, Peter Collins and Adam Smith, eds. Comparative Studies in Australian and New Zealand English: Grammar and Beyond. Amsterdam: Benjamins, 125-137.

Sayder, Stefan. 1989. The subjunctive in Indian, British and American English: A corpus-based study. Linguistische Arbeiten 69: 58-66.

Schneider, Edgar W. 2011. The subjunctive in Philippine English – An updated assessment. In M. L. S. Bautista, ed. Studies in Philippine English: Exploring the Philippine Component of the International Corpus of English. Mandaluyong City: Anvil Publishing, 159–173.

• • •

Ewa Jonsson

Mid-Sweden University and Uppsala University

Ewa.Jonsson@engelska.uu.se

Emotives in screen-mediated communication: from punctuation to emojis and beyond

Much of the punctuation in screen-mediated communication (via computers and cellphones) follows off-line patterns, albeit with more frequent repetition and omission of punctuation in the screen-mediated context. At the same time, screen-mediated communication is widely recognized for giving birth to the emoticon (e.g. the 'smiley,' *:), :-)*, the 'frownie,' *:-(*, and the 'winky,' *;-)*), a creative and innovative twist in the history of punctuation. This paper presents the results of a corpus-based, longitudinal

study of the distribution and function of emoticons and other typographical devices used to signal emotive cues in screen-mediated discourse. Following Jonsson (2015), emoticons are here regarded as instances of 'emotives,' as are certain emojis (minimal pictographs commonly realizing facial expressions, gestures and hearts).

Emotives are indicators of illocutionary force and chargers of evaluative, modal, attitudinal or affective meaning, typically found in written, screen-mediated texts produced for social interaction. Examples are emoticons, emojis, sentiment initialisms (e.g. *lol* for 'laughing out loud') and affection markers (such as *xx* for 'kisses'). They are highly context-dependent, indicating the tone in which a message might be interpreted, or internalized prosodically (cf. Knox 2009, Zappavigna 2012, Vandergriff 2013), thus sharing the pragmatic function of punctuation. In line with Knox (2009) and Zappavigna (2012), I argue that emotives are part of an ongoing evolution in punctuation, imposed by the immediacy of on-screen communication, by which punctuation today performs more interpersonal functions, evolving from its original textual, discourse-organizing function.

The qualitative discussion offered in the paper is supported by quantitative, empirical data from corpora of 21st century English online and cellphone communication. Two corpora of pre-emoji written screen-mediated communication in English are investigated: the UCOW computer chat corpus, collected in 2002 and 2004 (Jonsson 2015), and the CorTxt corpus of SMS text messages, collected in 2004–2007 (Tagg 2009, 2012). The distribution and functions of emotives in the corpora are compared to a Twitter sample corpus collected in 2016, containing emojis.

In my presentation, I will elucidate the status of emotives in language, their relationship to punctuation, and to what degree they can be regarded as paralinguistic, prolinguistic, extralinguistic or linguistic, by answering the following questions:

- What is the nature and distribution of emotives across the corpora?
- How were the functions of emojis achieved in pre-emoji written screen-mediated communication, if at all?
- What are the functions of emotives, and are there non-emotive emojis?
- What are the linguistic implications of emotives and some predictions for their future?

Among my findings is a trajectory whereby emoticons have gradually been supplemented, but not supplanted, by sentiment initialisms, affection markers, and recently an array of emojis, suggesting that messages are increasingly imbued with interpersonal cues, following Knox (2009:162) reasoning that '[T]he trajectory of interpersonal punctuation [...] begins with boundary marking, moves to punctuating speech function, and then to punctuating attitude and identity.'

In brief, the paper serves to shed light on contemporary developments in the pragmatics of punctuation, namely the evolvement of emotive, interpersonal functions.

References

CorTxt = A corpus of 11,036 SMS text messages. 2009. Compiled by C. Tagg.
Twitter sample corpus = A collection of 2,088 tweets sampled from the Twitter API on 2016-09-04. Collected by E. Jonsson.

UCOW = Uppsala Conversational Writing Corpus. 2004. Compiled by E. Jonsson.
Jonsson, E. (2015). Conversational writing: A multidimensional study of synchronous and supersynchronous computer-mediated communication. Frankfurt am Main: Peter Lang.
Knox, J. S. (2009). Punctuating the home page: image as language in an online newspaper. Discourse and Communication 3 (2), 145–72.
Tagg, C. (2009). A corpus linguistics study of SMS text messaging. PhD dissertation. Department of English, University of Birmingham.
Tagg, C. (2012). Discourse of text messaging: analysis of SMS communication. London: Continuum.
Vandergriff, I. (2013). Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. Journal of Pragmatics 51, 1–12.
Zappavigna, M. (2012). Discourse of Twitter and social media: How we use language to create affiliation in the web. London: Continuum.

• • •

Amelia Joulain-Jay
Lancaster University
a.t.joulain@lancaster.ac.uk

Describing collocation patterns in OCR data: are MI and LL reliable?

The increasing availability of digitized historical material opens up promising avenues of research for scholars in the Humanities. However, much of this material has been digitized using optical character recognition (OCR) procedures, which have reportedly low levels of accuracy when used on historical material (e.g. Holley 2009). For the British Library's c19th newspapers collection (see Conboy 2009) for example, Tanner et al. (2009, section 6) report an average word accuracy of 78%. Although OCR correction procedures exist, they do not achieve perfect results. For example, Overproof, a state-of-the-art commercial software, claims to 'reduce the number of articles missed by a keyword search due to OCR errors by over 50%' (OverProof 2014).

Since OCR errors affect word-counts, a key question for corpus linguists is whether common statistics can reliably be used with OCR data. This paper focuses on two statistics often used to describe collocation patterns: Mutual Information (MI) – an effect-size statistic often used to describe the strength of the relationship between two words – and Log Likelihood (LL) – a significance statistic often used to describe the amount of evidence underpinning a given collocation pattern.

To test the reliability of these two statistics in OCR data, I assembled a small corpus containing two corresponding datasets, one a subset of the OCR data constituting

the BL c19th newspapers collection (the uncorrected dataset), and the other, the re-OCR'ed and hand-corrected [by Erik Smitterberg, used with permission] counterpart to the first (the hand-corrected dataset). I then compared the Mutual Information and Log Likelihood statistics calculated for pairs of words in both datasets. To evaluate Overproof's corrections, I also produced a third dataset made up of the first dataset corrected by Overproof (the automatically-corrected dataset). Each dataset contains around 160,000 words.

I find that all OCR-derived MI statistics are within 5 points of difference from their gold counterpart. However, most are over-estimates, which is potentially problematic. MI rankings are broadly reliable for small spans, but become less reliable for larger spans. In addition, the use of an LL cut-off point increases the reliability of MI rankings. 90% of LL statistics are also within 5 points of difference from their gold counterpart, and most are unsurprisingly under-estimates. However, over-estimates also occur, which is potentially problematic. Using MI = 3 and LL = 10.83 as test cut-off points, both MI and LL attract high rates of false positives, a troubling result. These rates are higher when considering only words occurring at least 10 times, so using a frequency floor does not solve the problem. Correcting the OCR using Overproof makes an appreciable difference for both MI and LL statistics, with the most striking improvements obtained for larger spans. This suggests that researchers interested in exploring collocation patterns with large spans in OCR data would benefit from looking into OCR post-correction.

References

Conboy, M. (2009). The 19th Century British Library Newspapers Website. Reviews in History, (730). <http://www.history.ac.uk/reviews/print/review/730>, accessed 17/11/2014.

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, 15 (3/4).

OverProof (2014). OverProof: automatic correction of OCR. <http://overproof.projectcomputing.com/>, accessed 17/11/2014.

Tanner, S., Muñoz, T., & Pich Hemy, R. (2009). Measuring mass text digitization quality and usefulness. D-Lib Magazine, 15 (7/8), 1082-9873.

• • •

Mark Kaunisto
University of Tampere
mark.kaunisto@uta.fi

Juhani Rudanko
University of Tampere
juhani.rudanko@uta.fi

Agentivity and Complement Selection: a Case Study with Evidence from Large Corpora

Consider sentences (1a) and (1b), from COHA, the Corpus of Historical American English:

(1) a. … he was frightened to get out of bed in case that dreadful pain came roar ing back. (1998, FIC)
b. The family dog and cat also were frightened of leaving the house. (1964, NEWS)

In both (1a) and (1b) the matrix predicate is the adjective *frightened*, and it selects a non-finite sentential complement in both. In (1a) the complement is a *to* infinitive and in (1b) it is of the *of -ing* type, consisting of the preposition *of* and a gerundial clause. It is assumed, in accordance with traditional and current work, that in both (1a) and (1b) the sentential complement has an understood subject. In both sentences the matrix predicate assigns a semantic role to its subject, and both (1a) and (1b) are control constructions.

The two constructions selected by the adjective *frightened* and illustrated in (1a-b) are treated side by side in the entry for the adjective *frightened* for instance in the *OALD* (2005), suggesting that they are semantically close to each other. However, in current work on the syntax semantics interface it is generally assumed that a "difference in syntactic form always spells a difference in meaning" (Bolinger 1968, 127), and the paper compares *to* infinitives and *of -ing* complements on the basis of corpus evidence. Earlier work, including Allerton (1988) and Smith (2009), contains valuable insights, but the paper develops a new approach on the basis of the semantic role of the lower subject. In particular, the authors examine the hypothesis that agentive lower subjects are associated with *to* infinitive complements. This hypothesis has been put forward in very recent work on the basis of other higher predicates, and the present paper examines whether the hypothesis can shed light on the complement selection properties of the adjective *frightened*, not considered in earlier work. The broader task is to contribute to a fuller characterization of the semantic potential of each type of sentential complement in the system of English predicate complementation. Data from COHA show that the two constructions begin to occur side by side from about the 1950s and the 1960s onwards, and the study examines data from COCA,

the Corpus of Contemporary American English, to compare the semantics of the two non-finite patterns of sentential complements in very recent English.

References

Allerton, D. (1988). "Infinitivitis" in English. In J. Klegraf and D. Nehls (eds), Essays on the English Language and Applied Linguistics on the Occasion of Gerhard Nickel's 60th Birthday, 11-23. Heidelberg: Julius Groos.

Bolinger, D. (1968). Entailment and the Meaning of Structures. Glossa, 2, 199-127.

OALD = Oxford Advanced Learner's Dictionary of Current English, ed. by A. S. Hornby. (1979; 2005 7th edn.) Oxford: Oxford University Press.

Smith, M. (2009). The Semantics of Complementation in English: a Cognitive Semantic Account of Two English Complement Constructions. Language Sciences, 31, 360-388.

• • •

John Kirk
Dresden University of Technology / Belfast
jk@etinu.com

Vander Viana
University of Stirling
vander.viana@stir.ac.uk

How Homogenous are Academic Registers in ICE? A Corpus-driven Approach

The International Corpus of English continues to fulfill its aim of providing resources for comparative studies of the English used in countries where it is either a majority first language (for example, Canada and Australia) or an official additional language (for example, India and Nigeria). One part of its success has been the choice of comparable spoken and written text categories which comprise each ICE corpus and which provide the evidence for each national variety. Whereas there has been some global criticism of the ICE text category suitability for L2 countries (e.g. Leitner 1992, Schmied 1996), to the best of our knowledge, no-one has criticised the inclusion of academic writing in this corpus, which is divided in four subtypes (i.e. Humanities, Natural Sciences, Social Sciences, and Technology). At this ICAME conference, where the outcomes of a major review of the ICE project are to be discussed and agreed upon, it seems timely to take a critical look at this text category across the ICE corpora.

We will investigate the 16 available ICE corpora with the following research questions:

- What are the lexical characteristics of each ICE academic writing subtype?
- How homogeneous is the academic writing text category in English as a first/second language in ICE corpora?
- How convergent/divergent is academic prose across ICE corpora?

Our methodology is based on a key key word approach (Scott & Tribble 2006), which identifies words which are key (or used strikingly more or less than expected when compared to a reference corpus) in a number of different texts.

The results show considerable convergence as well as considerable divergence across national corpora. In the Humanities, for instance, we find that the subject matter varies great between Literature and the Arts, Society, Education, Philosophy, Politics or the discussion of languages. In Technology, we find that the subject matter shows concerns about methodology (what is involved in the research, and how the research is carried out), the environment and the natural world, and finance and economics.

The key key words in our study clearly show different ways of conceptualizing academic fields throughout the world. Whereas the results confirm preconceived notions, they raise new perspectives and concerns. The paper offers a fresh critique of ICE texts, and an addition to the study of corpus-driven academic prose (e.g. Scott & Tribble 2006, Biber 2006, Hyland 2009, Viana 2012) and the many recent studies in Hyland & Shaw (2016).

References

Biber, D. (2006). *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Hyland, K. (2009). *Academic Discourse in a Global Context*. London: Continuum.

Hyland, K. & Shaw, P. (Eds.). (2016). *The Routledge Handbook of English for Academic Purposes*. Abingdon: Routledge.

Leitner, G. (1992). International corpus of English: Corpus design – Problems and suggested solutions'. In G. Leitner (Ed.). *New Directions in English Language Corpora: Methodology, Results, Software Developments* (pp. 33–64). Berlin: Mouton de Gruyter.

Schmied, J. (1996). Second language corpora. In S. Greenbaum (Ed.), *Comparing English Worldwide: The International Corpus of English* (pp. 182–196). Oxford: Clarendon Press.

Scott, M. & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Viana, V. (2012). *Disciplinary Variation in Academic Writing: A Corpus Study of PhD Theses in English Language and Literature*. Unpublished PhD Thesis, Queen's University Belfast.

• • •

Haidee Kruger
Macquarie University / North-West University
haidee.kruger@mq.edu.au

Bertus Van Rooy
North-West University
bertus.vanrooy@nwu.ac.za

A multifactorial analysis of contact-induced constructional change in speech reporting in White South African English

In settings of prolonged language contact, language change may be the consequence of overt cross-linguistic influence (CLI) where one language adopts constructional forms from another, giving rise to a new form-meaning pair in the adopting language (Mougeon et al. 2005: 102). The process in which a new construction is created through CLI is classified as a form of "instantaneous constructionalisation" by Traugott and Trousdale (2013). Change may also result from covert CLI, where a shift in the frequency distribution of competing constructional forms occurs under influence of differences in the frequency of similar constructional forms in the contact language (Mougeon et al. 2005: 102-103). Covert CLI may initially only lead to constructional changes in the sense of Traugott and Trousdale (2013), or changes from minor to major use pattern in the terminology of Heine and Kuteva (2005), but in time, constructionalisation may result as the semantic or pragmatic associations of the forms change. Hilpert (2013: 16) subsumes all these processes in his definition of constructional change: "Constructional change selectively seizes a conventionalized form-meaning pair of a language, altering it in terms of its form, its function, any aspect of its frequency, its distribution in the linguistic community, or any combination of these."

In this paper we present a corpus analysis of constructional changes in the broader sense of Hilpert (2013) in a set of reported-speech constructions in White South African English (WSAfE), a native variety of English that has been in extensive contact with Afrikaans throughout its history. The analysis is based on comparable diachronic corpora of WSAfE (Wasserman & Van Rooy 2014), its parent variety British English (BrE) (represented by ARCHER), and the contact language, Afrikaans (Kirsten 2015). Three registers, fiction, newswriting and letters, are selected, and the timeframe is the 19th and 20th centuries (split in four half-century periods) for the two English corpora, but only the 20th century for Afrikaans. Three related sets of reported-speech constructions are analysed: (1) the position of the reporting clause in direct speech and thought, (2) quotative inversion in reporting clauses in non-initial position in direct and indirect speech and thought, and (3) the presence or absence of the complementiser *that* in indirect speech and thought. Features (1) and (3) represent potential instances of covert CLI, while feature (2) combines possibilities of overt and covert CLI. We undertake a multifactorial analysis of the effects of Variety, Register and Period on the selection of alternate constructions for each of the three features,

to retrace the course of constructional change associated with language contact in this particular setting.

The findings indicate divergent contact effects for the three features analysed. For feature (1), there is evidence of constructional change in an increasing preference for the final position of the reporting clause across both English varieties. This is particularly evident in newswriting, but with a steeper trajectory of change in WSAfE compared to ARCHER. In this case, an **existing change in progress in English is amplified by contact with Afrikaans**, also undergoing a similar change. For feature (2), overt transfer from Afrikaans does not take place. WSAfE follows the pattern of BrE in shifting from the inverted to non-inverted order. However, WSAfE lags behind BrE in the rate of change, which might be attributed to the preserving effect of Afrikaans inverted order. Here **contact therefore slows down change**. For feature (3) there is constructional change in the form of an increase in *that* omission over time in both English varieties, but at different rates in different registers. Published registers show more convergence between WSAfE and Afrikaans (which demonstrates a very high omission ratio) than letters, which are not subject to an editorial process, suggesting a role for **converging editorial norms** in the shared publishing context of WSAfE and Afrikaans.

References

Heine, B., & Kuteva, T. (2005). Language Contact and Grammatical Change. New York: Cambridge University Press.

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

Kirsten, J. (2015). The use of was in Afrikaans passive constructions: A diachronic corpus study. Southern African Linguistics and Applied Language Studies, 33 (2), 159-170.

Mougeon, R., Nadasdi, T. & Rehner, K. (2005). Contact-induced linguistic innovations on the continuum of language use: The case of French in Ontario. Bilingualism: Language and Cognition, 8 (2), 99-115.

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

Wasserman, R. & Van Rooy, B. (2014). The development of modals of obligation and necessity in White South African English through contact with Afrikaans. Journal of English Linguistics, 42 (1), 31-50.

• • •

Haidee Kruger
Macquarie University / North-West University
haidee.kruger@mq.edu.au

Adam Smith
Macquarie University
adam.smith@mq.edu.au

Bertus Van Rooy
North-West University
bertus.vanrooy@nwu.ac.za

Colloquialisation vs densification in the British and Australian Hansard: A diachronic multidimensional approach

Colloquialisation has been identified as an important process in recent short-term language change (Hundt & Mair 1999; Leech et al. 2009; Mair 2006). Colloquialisation can be defined as the drift of written registers towards more oral styles (Biber & Finegan 1989), where lexicogrammatical features associated with casual conversation spread to more formal written or spoken registers (Collins & Yao 2013: 480). This process has been documented across varieties of English, but Australian English is identified as leading changes in the direction of more colloquial usage (Collins 2013; Collins & Yao 2013), with other varieties, such as British English, comparatively conservative.

The analysis of colloquialisation is complicated by the fact that "there are facets of grammar where anti-colloquialization – a movement further away from spoken English norms – appears to be in the ascendant" (Leech et al. 2009: 245). This counterweight to colloquialisation is termed "densification", and involves "compacting meaning into a smaller number of words" (Leech et al. 2009: 249).

Previous studies of colloquialisation and densification across varieties of English are limited by the shortage of sufficient comparable diachronic data (see Collins & Yao in review). A further limitation is the tendency to focus on smaller sets of features investigated independently (Collins 2014, 2015; Peters 2014). This means that most current research in this area has given limited consideration to the internal relations among larger sets of potentially changing features and the interplay between colloquialisation and densification.

To address the first of these limitations, we have developed a comparable Australian and British Diachronic Hansard Corpus, covering the period 1901 to 2016. The Hansard, as edited written representation of speech (Slembrouck 1992), is a site where norms for speaking and writing, and informality and formality compete. This makes it particularly suitable for investigating the tensions between colloquialisation and densification.

We adapt the multidimensional analysis method of Biber (1988) to investigate lexicogrammatical changes from a diachronic perspective, since the method takes into statistical account the covariation of features. Firstly, we replicate Biber's (1988) mod-

el using the Hansard corpus, investigating how the language in the Hansard changes across the century in the two varieties, in relation to the dimensions identified by Biber (1988). We focus specifically on Dimensions 1, 3, 5 and 6, which relate most clearly to the distinction between informal and interactive spoken language, and formal and informational written language. Secondly, we analyse individual linguistic features from Biber's model that demonstrate a significant change in frequency across time. In this inductive analysis, we consider the direction of change as well as the pattern of change, grouping features together to develop a fine-grained interpretation of the co-occurrence patterns of features.

Preliminary findings indicate clear colloquialisation trends across a range of features, for example the decline of passives and the rise of *that* omission. Conversely, a decrease in verb frequency counterbalanced by an increase in the frequency of nouns indicates a move towards a more compressed, phrasal style. There is evidence of a complex interplay between these two trends over time across the two varieties.

References

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. & Finegan, E. (1989). Drift and evolution of English style: A history of three genres. Language, 65 (3), 487-517.

Collins, P. (2013). Grammatical variation in English worldwide: The role of colloquialization. Linguistics and the Human Sciences, 8 (3), 289-306.

Collins, P. (2014). Quasi-modals and modals in Australian English fiction 1800-1999, with comparisons across British and American English. Journal of English Linguistics, 42 (1), 7-30.

Collins, P. (2015). Diachronic variation in the grammar of Australian English: Corpus-based explorations. In P. Collins (Ed.), Grammatical Change in English World-Wide. Amsterdam: John Benjamins, 15-42.

Collins, P. & Yao, X. (In review). Colloquialisation in contemporary Australian English.

Collins, P. & Yao, X. (2013). Colloquial features in World Englishes. International Journal of Corpus Linguistics, 18 (4), 479-505.

Hundt, M. & Mair, C. (1999). "Agile" and "uptight" genres: The corpus-based approach to language change in progress. International Journal of Corpus Linguistics, 4 (2), 221-242.

Leech, G., Hundt, M., Mair, C. & Smith, N. (2009). Change in Contemporary English: A Grammatical Study. Cambridge: Cambridge University Press.

Mair, C. (2006). Twentieth-century English: History, Variation, and Standardization. Cambridge: Cambridge University Press.

Peters, P. (2014). Australian narrative voices and the colloquial element in nineteenth century written registers. Australian Journal of Linguistics, 34 (1), 100-117.

Slembrouck, S. (1992). The parliamentary Hansard "verbatim" report: The written construction of spoken discourse. Language and Literature 1 (2), 101-119.

• • •

Kristopher Kyle
University of Hawaii at Manoa
kkyle@hawaii.edu

Using corpus derived indices related to words, n-grams, and verb argument constructions to predict L2 writing quality

Usage-based perspectives on language development suggest that frequency is a key driver in language learning (Ellis 2002a, 2002b). Constructions that are more frequently encountered be learned earlier/more easily. Word frequency features are commonly used in language testing as a proxy for vocabulary knowledge in the evaluation of writing tasks (e.g., Jarvis 2002) and in the development of automatic essay scoring models (e.g., Guo, Crossley & McNamara, 2013). However, usage-based research extends beyond vocabulary items, and recent investigations have explored the relationship between holistic scores of writing quality and the corpus-based frequency of multi-word units (Bestgen & Granger 2014; Crossley, Cai, & McNamara 2012; Kyle & Crossley 2016, 2015). Kyle (2016), building on work by Ellis et al. (Ellis & Ferreira-Junior 2009; Römer, O'Donnell, & Ellis 2015) has also recently extended usage-based perspectives to the assessment of syntactic development through the use of verb argument construction (VAC) indices related to corpus frequency and strength of association. In this study, corpus derived indices related to three levels of linguistic abstraction (i.e., words, n-grams, and VACs) are used to model holistic scores of writing quality. The goals of this study are three-fold. First, the study examines the predictive validity of usage-based and corpus derived indices related to words, n-grams, and VACs. Second, the study examines the relative importance of indices related to words, n-grams, and VACs in explaining writing quality scores. Finally, by using only corpus derived indices, the study attempts to introduce an assessment methodology that can be extended to languages beyond English.

Method

The learner corpus used in the study comprised 480 independent essays written for the TOEFL iBT and scored by trained TOEFL raters. Indices related to word frequency and range, bigram and trigram frequency, range, and strength of association, and VAC frequency and strength of association were used to model essay quality scores. All indices were derived from the 90-million word academic section of the Corpus of Contemporary American English (COCA; Davies 2010) and are freely available in TAALES (Kyle & Crossley 2015) and TAASSC (Kyle 2016). After checking for statistical assumptions (e.g., normality and multicollinearity) a multiple regression analysis was conducted to model holistic scores of writing quality using the aforementioned variables.

Results & Discussion

The results of a multiple regression indicate that eight indices related to word frequency and range, trigram frequency and association strength, and VAC frequency and association strength explained 51.8% (r = .720, R2 = .518) of the variance in TOEFL essay quality scores. Respectively, indices related to words, n-grams and VACS ac-

counted for 42.4%, 6.7%, and 2.7% of the variance explained by the model. Higher scoring essays included word lemma types and VACs with lower reference corpus frequency, a higher proportion of frequent trigrams, and more strongly associated verb-VAC combinations and trigrams. In the presentation, the implications of these results will be discussed with regard to the predictive validity of corpus-derived indices, the relative importance of each index type, and the generalizability of the predictor model.

References

Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. Journal of Second Language Writing, 26, 28–41. http://doi.org/http://dx.doi.org/10.1016/j.jslw.2014.09.004

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, Paradigmatic, and Automatic N-Gram Approaches to Assessing Essay Quality. In Twenty-Fifth International FLAIRS Conference.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. Literary and Linguistic Computing, 25(4), 447–464. http://doi.org/10.1093/llc/fqq018

Ellis, N. C. (2002a). Frequency effects in language processing. Studies in Second Language Acquisition, 24(2), 143–188. Retrieved from http://dx.doi.org/10.1017/S0272263102002024

Ellis, N. C. (2002b). Reflections on frequency effects in language processing. Studies in Second Language Acquisition, 24(2), 297–339. Retrieved from http://dx.doi.org/10.1017/S0272263102002140

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. Annual Review of Cognitive Linguistics, 7(1), 188–221. http://doi.org/10.1075/arcl.7.08ell

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. Assessing Writing, 18(3), 218–238.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. Language Testing, 19(1), 57–84. http://doi.org/10.1191/0265532202lt220oa

Kyle, K. (2016). Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Georgia State University. Retrieved from http://scholarworks.gsu.edu/alesl_diss/35/

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. Journal of Second Language Writing, 34, 12–24. http://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. TESOL Quarterly, 49(4), 757–786. http://doi.org/10.1002/tesq.194

Römer, U., O'Donnell, M. B. & Ellis, N. C. (2015). Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions. In N. Groom, M. Charles & J. Suganthi (Eds.), Corpora, Grammar and Discourse: In honour of Susan Hunston. Amsterdam: John Benjamins Publishing Company.

• • •

Mikko Laitinen
University of Eastern Finland
mikko.laitinen@uef.fi

Placing ELF among the present-day varieties of English: Results from typological profiling

Previous corpus-based research of English as a lingua franca (ELF) has suggested that second language use and second language acquisition are essentially dissimilar. Mauranen (2012) notes that this fundamental fact leads to differences between ELF and learner Englishes. In addition, she points out that with regards to one genre of spoken academic ELF, "the overwhelming majority of lexis, phraseology, and structures are indistinguishable from those found in a comparable corpus of educated ENL, including their frequency distributions" (Mauranen 2012: 247). These observations are however based on a relatively small set of features.

This presentation puts these claims to a test and introduces a study that makes use of typological profiling based on aggregate structural criteria. This typological profiling method has previously been used to analyze a range of native Englishes, indigenized L2 varieties and learner English, but has not been used to assess ELF data. The method measures indices of grammatical analyticity, defined by the presence of free grammatical markers, and grammatical syntheticity, the presence of bound markers (e.g. Szmrecsanyi 2009; Szmrecsanyi & Kortmann 2011). Previous studies indicate substantial differences between various Englishes.

The study presented is corpus-based, and it draws evidence from the well-known first generation ELF corpora (viz. spoken VOICE by Seidlhofer (2011) and the new written WrELFA by Mauranen et al. (2015)). Additional evidence is drawn from the pilot versions of two second-generation ELF corpora that are not only larger in size in terms of the L1 backgrounds of the informants, but they also offer access to multi-genre evidence. The corpora target texts in which English is used an additional linguistic resource alongside people's L1s in Sweden and Finland, two countries in which the role of English is undergoing changes whereby it is increasingly being adopted as additional linguistic resource alongside the main domestic languages in the two countries (cf. Taavitsainen & Pahta 2008; Leppänen et al. 2011; Bolton & Meierkord 2013).

The results will shed light on the typological status of ELF and provide empirical evidence of its structural characteristics in a range of genres. The results are compared with the observations in Szmrecsanyi & Kortmann (2011) in which learner language is characterized by overuse of analytic markers and underuse of synthetic markers relative to the main Standard English varieties. The quantitative results contain very few traces of this quantitative tendency in lingua franca data, and they therefore have considerable implications of how ELF is conceptualized in English corpus linguistics. In addition, the results support Mauranen's (2012) arguments and show close similarities on purely structural grounds between the various genres in ELF and many genres in the main Standard English varieties.

References

Bolton, K. & Meierkord C. (2013). English in contemporary Sweden: Perceptions, policies, and narrated practices. Journal of Sociolinguistics 17(1), 93–117.

Leppänen, S. et al. (2011). National Survey on the English Language in Finland: Uses, Meanings and Attitudes. http://www.helsinki.fi/varieng/journal/volumes/05/. (accessed 10 February 2017).

Mauranen, A. (2012). Exploring ELF: Academic English Shaped by Non-Native Speakers. Cambridge: Cambridge University Press.

Mauranen, A., Carey R. & Ranta E. (2015). New answers to familiar questions: English as a lingua franca. In D. Biber & R. Reppen (eds.), Cambridge Handbook of English Corpus Linguistics. Cambridge: Cambridge University Press, 401–417.

Seidlhofer, B. (2011). Understanding English as a Lingua Franca. Oxford: Oxford University Press.

Szmrecsanyi, B. (2009). Typological parameters of interlingual variability: Grammatical analyticity vs. syntheticity in varieties of English. Language Variation and Change 21(3), 319–353.

Szmrecsanyi, B. & Kortmann B. (2011). Typological profiling: Learner Englishes versus indigenized L2 varieties of English. In J. Mukherjee & M. Hundt (eds.), Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap. Amsterdam: John Benjamins, 167–187.

Taavitsainen, I. & Pahta, P. (2008). 'From global language use to local meanings: English in Finnish public discourse.' English Today, 24(3), 25–38.

• • •

Alexander Lakaw
Linnaeus University
alexander.lakaw@lnu.se

Diachronic Shifts in Agreement Patterns of Collective Nouns in American and British English in the 19th and early 20th century

Variation in agreement with English collective nouns (as illustrated in (1) and (2)) has received a great deal of attention in corpus linguistics.

(1) The *army was* not in winter quarters now; *it* was in the field fighting, (COHA, 1913)
(2) the *army have* gone into winter quarters (COHA, 1823)

Previous synchronic research (e.g., Levin 2001; Hundt 2006) has shown that present-day AmE prefers singular agreement to a greater extent than BrE. However, much less is known about the varying agreement patterns in the LME period, in which collective nouns were "notoriously troublesome as to number, and there has been much fluctuation over time" (Denison 1998: 99), and so far it has not been established when the two varieties started diverging in their preferences. This diachronic study compares the agreement patterns in AmE and BrE and draws evidence from the Corpus of Historical American English (COHA), the Old Bailey Corpus (OBC) and the Corpus of Late Modern English Texts (CLMET). As a complement to the synchronic investigations of this phenomenon, and in an effort to understand the emerging differences between the varieties, this study covers the understudied time span 1810–1949 and includes the agreement patterns of a range of collective nouns from six semantic categories, which are (1) EMPLOYEES (e.g., *crew*), (2) FAMILY (e.g., *couple*), (3) MILITARY (e.g., *army*), (4) POLITICS (e.g., *government*), (5) PUBLIC ORDER (e.g., *police*) and (6) SOCIETY (e.g., *generation*).

The results show an overall increase of singular agreement in both varieties during the 19th century. Moreover, the findings suggest that verbal and pronominal agreement patterns behave differently in that the latter is more likely to be plural, and also that variation exists amongst the semantic categories. Even though a more in-depth genre analysis is beyond the scope of this study due to the diversity of the corpora used, preliminary results for AmE indicate that plural agreement is most frequent in fiction (probably due to the number of spoken-like features), whereas the vast majority of singular verbs and pronouns occurred non-fiction genres, e.g. magazines.

Finally, the main finding of the study is that the incipient stages of a change in AmE towards a preference for singular agreement are visible in the 19th-century material, but the expected leading role of AmE in this change (cf. Collins 2015: 29) could not be established in the corpora used. Instead, the analysis of the 20th-century material suggests that AmE displays signs of a kick-down development (Hundt 2009: 33) in which BrE shows a greater tendency for the singular in the 19th century, but is overtaken by AmE in the early 20th century, resulting in the singular being the

dominant choice of agreement in that variety by the 1940s. This late but fast developing change in AmE might be explained by extra-linguistic influences, as for instance prescriptivism and the incipient Standard AmE as a focused variety. Considering the historical context of the time, this result contributes to the study of the emergence of AmE as a main variety of English since "[j]udging American English within the context of a British norm continued until the conclusion of World War II, when the United States emerged as a world power, giving American English greater international prominence" (Kretschmar & Meyer 2012: 141).

References

Collins, P. (2015). Diachronic variation in the grammar of Australian English. In P. Collins (ed.), Grammatical Change in English World-Wide. John Benjamins, 15–42.

Denison, D. (1998). Syntax. In S. Romaine (ed.), The Cambridge history of the English language: 1776–1997, Vol. 4. Cambridge: Cambridge University Press, 92–329.

Hundt, M. (2006). The committee has/have decided …On concord patterns with collective nouns in inner and outer circle varieties of English. Journal of English Linguistics, 34 (3), 206–232.

Hundt, M. (2009). Colonial lag, colonial innovation, or simply language change? In G. Rohdenburg & J. Schlüter (eds.), One Language, Two Grammars. Cambridge: Cambridge University Press, 13–37.

Kretschmar, W. A. & Meyer, C. F. (2012). The idea of Standard American English. In R. Hickey (ed.), Standards of English. Codified varieties around the world. Cambridge: Cambridge University Press, 139–158.

Levin, M. (2001). Agreement with Collective Nouns in English. Stockholm: Almquist and Wiksell.

• • •

Tove Larsson
Uppsala University
tove.larsson@engelska.uu.se

Henrik Kaatari
Uppsala University
henrik.kaatari@engelska.uu.se

Extraposition in learner and expert writing: Rethinking informality

Subject extraposition, as in *it is interesting to note the difference*, is a common feature of formal writing (Biber et al. 1999:722). This paper aims to investigate to what extent genre differences can be used to explain seemingly (in)formal uses of this construction with a *to*-infinitive clause in learner and expert writing, as outlined below.

The construction includes an anticipatory *it* and an extraposed clausal subject (Quirk et al. 1985:1391). It enables writers to depersonalize claims (Kaltenböck 2005), and it has been found to be especially common in formal, academic discourse (Kaatari 2016). Since many studies have noted that learners tend not to achieve an appropriate level of formality in their writing (Altenberg & Tapper 1998; Larsson, forthcoming), one could expect learners to struggle with this construction. Indeed, studies such as Hewings & Hewings (2002) and Römer (2009) have found indications of informal use; however, the claims made in these studies are mainly based on lexical misuse (e.g. *it is amazing that*) or under/overuse of the construction, and few additional steps have been taken towards investigating informal use more systematically.

The present study goes beyond previous research by carrying out a detailed investigation of high-frequency realizations of the construction that have been found to vary across genres (cf. Groom 2005) to test claims of (in)formality. More specifically, the study focuses on the most frequent predicate and clausal type of the construction, namely adjective phrases followed by *to*-infinitive clauses (e.g. *it is* **important** *to* **remember**; *it is* **fair** *to* **say**). It compares learner writing (L1 Swedish) from one genre (academic writing) to expert writing from several different genres (academic writing, popular science, news, novels and conversation). In doing so, unlike the above-mentioned previous studies, this study investigates (i) which of the experts' genres the learners' use is closest to and, thus, (ii) whether describing learner use as simply "informal" might be simplifying matters.

The study uses data from two corpora, ALEC and BNC-15. ALEC (the *Advanced Learner English Corpus*) is a recently compiled 1.3-million-word corpus of learner academic writing. BNC-15 is a methodically sampled 3-million-word subset of the BNC that is constructed to enable a variety of different genre comparisons. The study uses inferential statistics such as *co-varying collexeme analysis* (Stefanowitsch & Gries, 2005). Preliminary results show that while the learners use some of the adjective-verb pairings that are associated with academic writing in the expert data (e.g. *it is* **reasonable** *to* **assume**), some pairings seem to resemble the use in the other expert genres

more closely than that of the experts' academic texts. For example, the learners tend to use pairings that rank highly in the non-academic genres of the expert data with regard to collostruction strength, such as *it is* **easy**/**hard** *to* **get** and *it is* **interesting** *to* **see**. It is hoped that the findings of the study will contribute to a more nuanced view of (in)formal uses in learner data, thus benefitting both L2 instruction and theory.

References

Advanced Learner English Corpus (ALEC). Corpus compiled at Uppsala University in 2013.

Altenberg, B. & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), Learner English on computer. London: Longman, 80–93.

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

BNC-15. Subset of the British National Corpus (BNC) sampled in 2012.

Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. Journal of English for Academic Purposes, 4 (3), 257–277.

Hewings, M. & Hewings, A. (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.

Kaatari, H. (2016). Variation across two dimensions: Testing the Complexity Principle and the Uniform Information Density Principle on adjectival data. English Language and Linguistics, 20 (3), 533–558.

Kaltenböck, G. (2005). It-extraposition in English: A functional view. International Journal of Corpus Linguistics, 10 (2), 119–159.

Larsson, T. (forthcoming). A syntactic analysis of the introductory it pattern in non-native-speaker and native-speaker student writing. In M. Mahlberg & V. Wiegand (Eds.), Corpus linguistics, context and culture. Berlin: De Gruyter Mouton.

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

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. Annual Review of Cognitive Linguistics, 7 (1), 140–162.

Stefanowitsch, A. & Gries, S. Th. (2005). Co-varying collexemes. Corpus Linguistics and Linguistic Theory, 1 (1), 1–43.

• • •

Aatu Liimatta
University of Helsinki
aatu.liimatta@helsinki.fi

Exploring register variation on Reddit: a multi-dimensional study of language use on a social media website

Social media platforms are increasingly important forums for conversation. While some social media, such as Twitter, have received considerable attention from linguists (see e.g. Huang et al. 2015), many others have been relatively unstudied. Moreover, studies on register variation within online communities are particularly few and far between.

This paper focuses on Reddit, the third most popular English-speaking social media website (after Facebook and Twitter). Reddit is made up of thousands of „subreddits", user-created sub-communities centered around various topics, ranging from very general to very specific (e.g. news, politics, technology, games, entertainment, in-jokes and memes). Reddit users (or redditors) subscribe to and make posts in subreddits which interest them. The comment sections of these posts are often home to lively discussions, which encompass various registers from an academic style to casual conversation.

In this paper, I explore register variation within Reddit using Biber's multi-dimensional analysis (Biber 1988). The analysis extracts dimensions of register variation from texts based on a statistical analysis of the relative frequencies of dozens of linguistic features. These dimensions are then labeled according to the situational concerns from which they arise.

The data analyzed in this paper covers over 10,000 Reddit threads, comprising over 17 million words, posted in 37 subreddits during a period of one month in July 2015. From these threads, three register dimensions can be extracted. I compare the positions of different subreddits along these dimensions, and show how the dimensions reflect the various ways in which different situational concerns affect the produced text.

The most important of the three dimensions, „Personal vs. Factual Focus", relates to how language is used differently for talking about people and subjective issues (such as personal views and opinions) on one hand and for presentation of factual matter on the other. The „Informational vs. Involved Style" dimension distinguishes between an informationally more dense style of writing and a more casual, involved, personal style; this dimension is functionally very similar to Biber's (1988) first dimension. The „Non-Past vs. Past Focus" dimension describes the temporal focus of the register. These three dimensions are also more or less in line with the three „universals of register variation" suggested by Biber (2014).

I also discuss the overall suitability of the multi-dimensional method for analysing the asynchronous online discussions on Reddit (and other similar online contexts), and identify some issues to be tackled in future studies on register variation on Reddit and other social media.

References

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

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

Huang, Y., Guo, D., Kasakoff, A. & Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. Computers, Environment and Urban Systems, 59, 244-255.

• • •

María José López-Couso
University of Santiago de Compostela
mjlopez.couso@usc.es

Belén Méndez-Naya
University of Santiago de Compostela
belen.mendez@usc.es

From happenstance to epistemic possibility: Corpus evidence for the adverbialization of happenstance expressions

Modality is a wide semantic domain with various manifestations, ranging from grammatical devices (e.g. synthetic and analytic mood systems) to those closest to the lexical pole, as is the case of modal adverbs (see Huddleston & Pullum et al. 2002: 173). Within this broad domain, our ongoing research project is concerned with the expression of epistemic possibility, in particular with the origin, development, and present-day use of a number of modal adverbs and expressions, such as *it seems* (López-Couso & Méndez-Naya 2014a, 2014b), *looks like* (López-Couso & Méndez-Naya 2014c), and *maybe* (López-Couso & Méndez-Naya 2016), which convey some degree of doubt towards the truth of the speaker's proposition (Quirk et al. 1985: 620).

A common cross-linguistic source for expressions of epistemic possibility is happenstance or contingence, i.e. something that happens (by chance). Examples include Sanskrit *sam-bhavá* (‚occurrence' > ‚capacity, ability, possibility'; Monier-Williams 1899: 1179), Latin *forte* (ablative of *fors* ‚fortune, chance'), *forsitan* (< *fors sīt an* ‚chance be that'), and *fortasse* (< **forte an sīt/s*) (de Vaan 2008: 236), and Dutch *misschien* (< *tmachschien* ‚it may happen'; Beijering 2010: 5). Surprisingly, the concept of happenstance as the source for possibility expressions in English has (to our knowledge) not received due attention in the literature, despite the fact that it lies at the origin of two frequent epistemic adverbs, namely *perhaps* and *maybe* (Biber et al. 1999: 869). In a previous study based on dictionary evidence (OED, MED, and HTOED) we

drew attention to this rather underexplored area (López-Couso & Méndez-Naya 2015), and showed that *perhaps* and *maybe* are just two members of an extensive inventory of happenstance adverbial expressions available in Late Middle English and Early Modern English, which also included low-frequency formations such as *peraventure*, *by hap*, *may fall*, *may fortune*, and *chance*, among others. The present paper serves as a timely complement to our earlier work by providing empirical corpus evidence for the development and distribution of this set of adverbial elements over time. More specifically, our aims in this presentation are:

(i) to trace the developmental pathways followed by these expressions: from phrase to adverb (e.g. *perhaps* < Latin preposition *per* ‚for, by' + Scandinavian noun *hap* ‚occurrence, chance' + plural/adverbial *-s*) and from clause to adverb (e.g. *maybe* < *(it) may be that…*), and approach their process of adverbialization from the perspective of grammaticalization and (inter)subjectification;
(ii) to examine the distribution of the various members of the catalogue of happenstance expressions across time and register; and
(iii) to identify the different syntactic functions realized by these adverbial elements and the positions they occupy in the clause.

For our purposes, data have been drawn from various sources: on the one hand, the Penn Historical Corpora, which allow us to look into the long diachrony of happenstance epistemic expressions; on the other, the 525-million-word corpus EEBOCorp 1.0 (Petré 2013), covering the period 1474-1700, which proves a particularly suitable source of evidence to zoom in on the critical period in their historical development.

References

Beijering, K. (2010). The grammaticalization of Mainland Scandinavian MAYBE. Bergen Language and Linguistics Studies 1. DOI: http://dx.doi.org/10.15845/bells.v1i1.39

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). Longman Grammar of Spoken and Written English. London: Longman.

de Vaan, M. (2008). The Etymological Dictionary of Latin and the Other Italic Languages. Leiden and Boston: Brill.

HTOED = Historical Thesaurus of the Oxford English Dictionary <http://public.oed.com/historical-thesaurus-of-the-oed/> Ed. by C. Kay, J. Roberts, M. Samuels & I. Wotherspoon.

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

Kroch, A. & Taylor, A. (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4. <http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4>

Kroch, A., Santorini, B. & Diertani, A. (2016). The Penn Parsed Corpus of Modern British English (PPCMBE2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 1. <http://www.ling.upenn.edu/ppche-release-2016/PPCMBE2-RELEASE-1>

Kroch, A., Santorini, B. & Delfs, L. (2004). The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3. <http://www.ling.upenn.edu/ppche-release-2016/PPCEME-RELEASE-3>

López-Couso, M. J. & Méndez-Naya, B. (2014a). From clause to pragmatic marker: A study of the development of like-parentheticals in American English. Journal of Historical Pragmatics 15 (1), 66-91.

López-Couso, M. J. & Méndez-Naya, B. (2014b). On the origin of clausal parenthetical constructions: Epistemic/evidential parentheticals with seem and impersonal think. In I. Taavitsainen, A. H. Jucker & J. Tuominen (Eds.), Diachronic Corpus Pragmatics. Amsterdam & Philadelphia: John Benjamins, 189-212.

López-Couso, M. J. & Méndez-Naya, B. (2014c). Epistemic parentheticals with seem: Late Modern English in focus. In M. Hundt (Ed.), The Syntax of Late Modern English. Cambridge: Cambridge University Press, 291-308.

López-Couso, M. J. & Méndez-Naya, B. (2015). On the haps and mishaps of happenstance expressions as a source of epistemic adverbs in English. Paper presented at SHEL-9. Vancouver, June 2015.

López-Couso, M. J. & Méndez-Naya, B. (2016). From clause to adverb: On the history of maybe. In G. Kaltenböck, E. Keizer & A. Lohmann (Eds.) Outside the Clause. Amsterdam: John Benjamins, 157-176.

MED = Kurath, H. et al. (1952-2001). Middle English Dictionary. Ann Arbor: University of Michigan Press. <http://ets.umdl.mich.edu/m/med/>

Monier-Williams, M. (1899). A Sanskrit-English Dictionary: Etymologically and Philologically Arranged with Special Reference to Cognate Indo-European Languages. Revised by E. Leumann, C. Cappeller et al. Oxford: Clarendon Press.

OED = Oxford English Dictionary Online. Oxford University Press. <http://www.oed.com>

Petré, P. (2013). EEBOCorp 1.0. oai:lirias.kuleuven.be:123456789/416330.

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

• • •

Robbie Love
Lancaster University
r.m.love@lancaster.ac.uk

FUCK in spoken British English revisited with the Spoken BNC2014

This paper reports on an analysis of the word *fuck* in a sample of the Spoken British National Corpus 2014 (Spoken BNC2014; Love et al. 2017 fc), comparing this corpus with the original Spoken British National Corpus (Spoken BNC1994; Leech 1993), with a view to then extending this analysis to a larger set of bad language words (BLWs, McEnery 2005). The Spoken BNC2014 comprises transcripts of spontaneous, present day, informal conversations between speakers of British English which is being compiled by Lancaster University and Cambridge University Press and amounts to eleven million words (the sample used in this paper contains only the data available at the time of the study, which amounts to five million words).

This study is informed by the work of McEnery & Xiao (2004), who analysed the sociolinguistic distribution of the BLW FUCK in the ten million word spoken component of the Spoken BNC1994. They found that, in early 1990s spoken British English, when all morphological forms of FUCK are taken together:

- It is used by male speakers more than twice as frequently as female speakers (p. 240).
- Teenagers, children and young adults (in that order of rank) have a high propensity to use FUCK; however it is used by speakers in their late forties/fifties unexpectedly more so than speakers in their late thirties/early forties (p. 242).
- It is used most by speakers in social classes C2 and DE, but more so by AB than C1 (p. 243).

By replicating the approaches of McEnery & Xiao (2004) on the new data, the aims of this study are (a) to assess the extent to which the use and distribution of FUCK has changed in spoken British English in the last two decades, according to the sociolinguistic variables of *gender, age* and *socio-economic status*, and (b) to reflect on and respond to the methodological challenges presented by this sort of work in preparation for analysing many more bad language words.

Preliminary results suggest that in present day spoken British English:

- FUCK is now used equally as frequently by male and female speakers.
- The use of FUCK peaks among speakers in their twenties and decreases with age, apart from the 60-69 group which has a higher frequency than 50-59.
- The distribution of FUCK according to social class is similar to that of McEnery & Xiao (2004) but only if the same classification scheme (Social Grade) is used. If a newer scheme is used (NS-SEC), then it is speakers in the middle of the scale that seem to use fuck the most rather than those towards the bottom.

Clearly the most radical change is the levelling off of the previously attested gender gap; perhaps suggestive of the erosion of stereotypical divisions between 'male' and 'female speech'. With regards to age, the same pattern as observed by McEnery & Xiao (2004) occurs, but a decade later; perhaps this is reflective of people become parents later in their lives than they did two decades ago, and therefore being less likely to use bad language around their children.

The compilation of the Spoken BNC2014 has facilitated large-scale, diachronic analyses of spoken data on a scale which has until now not been possible. This study therefore exemplifies new challenges in the sociolinguistic study of spoken data. Using bad language as an appropriate case study I draw attention to some of the methodological challenges of comparing such datasets for sociolinguistic purposes and the implications such challenges may have for sociolinguistic generalisations (cf. Brezina & Meyerhoff 2014). The next step for this work is (a) to repeat this analysis for a large set of BLWs in the Spoken BNC2014 and (b) to analyse the BLWs qualitatively using McEnery's bad language categorization scheme (2005: 27).

References

Brezina, V. & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. International Journal of Corpus Linguistics, 19 (1), 1-28.

Leech, G. (1993). 100 million words of English. English Today, 9-15.

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

McEnery, T. (2005). Swearing in English: Bad language, purity and power from 1586 to the present. New York: Routledge.

McEnery, T. & Xiao, Z. (2004). Swearing in modern British English: the case of FUCK in the BNC. Language and Literature, 13 (3), 235-268.

• • •

Ursula Lutzky
Vienna University of Economics and Business
ursula.lutzky@wu.ac.at

You can ask us on Twitter. Twitter as a means of customer communication for British train companies

During the last decade, the microblogging platform Twitter has increasingly been the focus of academic research. Several studies have explored, among others, the use of Twitter in the context of politics (e.g. Mirer & Bode 2015, Zappavigna 2012), as a means of crisis communication (e.g. Bruns & Burgess 2014), or for journalistic purposes (e.g. Barnard 2016, Parmelee 2013). Thus, Twitter has been studied widely as a reporting tool used to tell and share stories in response to the tagline "What's happening?" (see also Page 2012, Papacharissi 2016).

One area of application that has not been researched extensively to date is businesses' use of Twitter as a channel for customer communication (but see Page 2014). This channel is used increasingly by British train companies, most of which have a Twitter account and encourage customers to contact them through this medium if they need "information or assistance related to [their] travel" (see https://www.eastmidlandstrains.co.uk/information/contact-us/). Indeed, customers are assured that they can tweet train companies 24/7 while being promised the most recent updates by following them on Twitter.

This study takes a closer look at the interaction between social media advisors employed by British train companies and their customers on Twitter. To this end, a corpus of tweets was compiled between the 1st and the 31st of August 2016 which comprises a total of 4.2 million words (including retweets). It is based on 37 different companies providing inter-city and regional train services in the United Kingdom and includes tweets by customers directed at these companies and their replies. Consequently, this corpus allows for the interactive potential of the medium to be explored in the context of business communication.

The analysis will initially provide a general assessment of language use in this corpus, including an account of Twitter specific features such as hashtags (e.g. *#virgin*, *#Cambridge*, *#useless*, *#badservice*), to then move on to a detailed discussion of frequent clusters and expressive speech acts (e.g. *sorry for the delay*, *thanks for letting us know*, *to make a complaint please*). In addition to quantitative findings, this paper will present specific examples of interactions between customers and social media advisors to take a closer look at consecutive exchanges on a specific topic from a pragmatic perspective. Thus, this study aims to gain further insights into the features of customer communication on social media through the analysis of a customised Twitter corpus and it will thereby contribute to the theme of the conference by interpreting one growing component of train companies' customer services through corpus linguistic means.

References

Barnard, S. R. (2016). 'Tweet or be sacked': Twitter and the new elements of journalistic practice. Journalism, 17 (2), 190-207.

Bruns, A. & Burgess, J. (2014). Crisis communication in natural disasters: The Queensland floods and Christchurch earthquakes. In K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann (Eds.), Twitter and Society. New York: Peter Lang, 373-384.

Mirer, M. L. & Bode, L. (2015). Tweeting in defeat: How candidates concede and claim victory in 140 characters. New Media & Society, 17 (3), 453-469.

Page, R. (2012). Stories and Social Media. London: Routledge.

Page, R. (2014). Saying 'sorry': Corporate apologies posted on Twitter. Journal of Pragmatics, 62, 30-45.

Papacharissi, Z. (2016). Affective publics and structures of storytelling: sentiment, events and mediality. Information, Communication & Society, 19 (3), 307-324.

Parmelee, J. H. (2013). Political journalists and Twitter: Influences on norms and practices. Journal of Media Practice, 14 (4), 291-305.

Zappavigna, M. (2012). Discourse of Twitter and Social Media. London: Bloomsbury.

• • •

Seth Mehl
University of Sheffield
s.mehl@sheffield.ac.uk

Susan Fitzmaurice
University of Sheffield
s.fitzmaurice@sheffield.ac.uk

Innovations in measuring lexical co-occurrence: Improved implementation for the automatic analysis of discursive meaning

Linguistic DNA is a major research project mapping semantic and conceptual change in Early Modern English, whose data-driven approach is built on computational analysis of every word in every text of Early English Books Online (specifically EEBO-TCP). The data set contains 55,000 printed Early English texts, and over one billion words, hand-keyed by the Text Creation Partnership, and including all of the spelling variation and printing errors and idiosyncrasies common to Early Modern English. This paper describes the innovative methods employed in automatically analysing discursive meanings in that very large data set. The project begins by employing MorphAdorner (Burns 2013) for tokenising and lemmatising the data. It then analy-

ses Pointwise Mutual Information (PMI) for co-occurring word lemmas, by creating symmetrical co-occurrence matrices representing all lemmas. PMI analysis as applied here is novel in three key ways, each of which is discussed in detail. First, PMI is measured for co-occurring lemmas across very large proximity windows of +/-50 words and +/-100 words, reflecting the project's theoretical framework for linguistic meaning. That is, the object of study here is meaning as it is constructed in discursive contexts well beyond the level of the utterance or sentence. The present approach is therefore very different from traditional collocational studies, which often measured co-occurrence within proximity windows of approximately five words, and is also different from much recent distributional semantic work, which often measures co-occurrence within windows of up to 10 words (cf. Burgess and Lund 1997). The present approach is closer to that of Landaure and Dumais (1997), who measure co-occurrence within paragraphs – with the caveats that EEBO-TCP is not consistently coded for paragraphs, and that the nature and purpose of paragraphs has evolved dramatically since Early Modern times. Second, PMI is measured against a grammatically defined statistical baseline rather than a traditional baseline of all words in the data set, in line with Bowie et al. (2013) and in contrast to CQPWeb (Hardie 2012), among others. Such a baseline reflects linguistic probability more meaningfully by dramatically reducing invariant Type C terms (cf. Wallis 2014). Finally, the project applies an innovative technique to identify not just pairs of co-occurring words, but also trios and quartets of co-occurring words within the given proximity windows. This technique can be conceptualised as co-occurrence matrices in which rows represent all co-occurring lemma pairs, and columns represent all lemmas, such that co-occurring trios of lemmas can be identified, and PMI scores can be calculated for those trios. Subsequent matrices expand the technique to measure PMI for quartets and larger word sets.

References

Burgess, C. & Lund, K. (1997). Modelling Parsing Constraints with High-dimensional Context Space. Language and Cognitive Processes, 12 (2/3), 177–210.

Burns, P. R. (2013). MorphAdorner v2: A Java library for the morphological adornment of English language texts. Evanston, IL: Northwestern University. https://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf. Accessed 29 November, 2016.

Hardie, A. (2012). CQPWeb – combining power, flexibility and usability in a corpus analysis tool. International journal of corpus linguistics, 17 (3), 380-409.

Wallis, S. (2014). Is language really a set of alternations? In corplingstats. https://corplingstats.wordpress.com/2014/02/20/is-language-alternations/. Accessed 1 December, 2014.

Bowie, J., Wallis, S. & Aarts, B.. (2013). The perfect in spoken British English. In B. Aarts, J. Close, G. Leech & S. Wallis (Eds.), The verb phrase in English: Investigating recent language change with corpora. Cambridge: Cambridge University Press, 318-52.

• • •

Seth Mehl
University of Sheffield
s.mehl@sheffield.ac.uk

Corpus frequency and cognitive salience: Finding correlations

Gilquin (2008) reported that light senses of verbs (e.g. *make a contribution*, *give support*) tend to outnumber concrete senses of those same verbs (e.g. *make furniture*, *give books*) in naturally occurring corpus data, whereas concrete senses tend to outnumber light senses in responses to elicitation tests. To make this point, Gilquin (2008) conducted a semasiological corpus study of the Switchboard and FROWN corpora, and also conducted semasiological elicitation tests with native speakers, in which respondents were asked to generate the first sentence that came to mind with the target verbs *give* and *take*. Affirming the importance of Gilquin's study, Werner and Mukherjee (2012) replicated the corpus portion of the study using selected International Corpus of English texts.

The differences between corpus frequency and cognitive salience remain an important and much-discussed question (cf. Glynn 2014, Arppe et al. 2010, Nordquist 2009, Gilquin 2006). The question is particularly complicated because both *corpus frequency* and *cognitive salience* are difficult to define, and are often left undefined. Corpus linguists do not often explicitly explore definitions of *frequency*, and explicit discussion tends to focus on the relationship between raw counts and normalisation per million words (cf. Lindquist 2009: 41-2, McEnery et al. 2006: 52-3, McEnery and Wilson 2001: 83, Evison 2010: 126). *Cognitive salience* is defined in many different ways, including traditional tests for psychological categories and reference points (cf. Rosch 1975a, 1975b, 1973), as well as measures of corpus frequency (cf. Arppe et al. 2010, Gilquin 2008, Heylen *et al.* 2008, Geeraerts 2006 [1989], Gilquin 2006, Gries 2006). *Cognitive salience* relates to various differing theories of prototypicality and entrenchment (cf. Taylor 2012, 2003; Schmid 2007, Geeraerts 1988).

In corpus semantics, frequency can reasonably be normalised semasiologically or onomasiologically (cf. Glynn 2014, Geeraerts 1997), rather than per million words. Fundamental methods of measuring corpus frequencies are the question at the heart of the present paper, which poses the following research questions:

1. Are Gilquin's (2008) semasiological measures of light senses and concrete senses corroborated by new semasiological observations of concrete and light senses of *make*, *take*, and *give* in speech and writing in the International Corpus of English component representing Great Britain (ICE-GB)?
2. Do corpus frequencies of light senses and concrete senses in ICE-GB actually differ from published elicitation test results, if the corpus frequencies are measured onomasiologically rather than semasiologically?

I present both semasiological and onomasiological analyses of the concrete and light senses of *make*, *take*, and *give* in ICE-GB. Manual semantic analysis is performed on nearly six thousand instances of *make*, *take*, and *give*, along with thousands of instances of their onomasiological alternates. Findings indicate that corpus frequen-

cies in speech (but not writing) do in fact correlate with elicitation test results, if the corpus frequencies are measured onomasiologically rather than semasiologically. I refer to Geeraerts's (2010) hypothesis of *onomasiological salience* in explaining this correlation, and compare that hypothesis to other hypotheses regarding corpus frequency, salience, and entrenchment.

References

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M. & Zeschel, A.. (2010). Cognitive corpus linguistics: Five points of debate on current theory and methodology. Corpora, 5 (1), 1-27.

Evison, J. (2010). What are the basics of analysing a corpus? In A. O'Keefe & M. McCarthy (Eds.), The Routledge handbook of corpus linguistics. London: Routledge, 122-35.

Geeraerts, D. (1988). Where does prototypicality come from? In B. Rudzka-Ostyn (Ed.), Topics in Cognitive Linguistics. Amsterdam: John Benjamins, 207-29.

Geeraerts, D. (1997). Diachronic prototype semantics: A contribution to historical lexicology. Oxford: Clarendon Press.

Geeraerts, D. (2006 [1989]). Prospects and problems of prototype theory. In D. Geeraerts, Words and other wonders. Berlin: Mouton de Gruyter, 3-26.

Geeraerts, D. (2010). Theories of lexical semantics. Oxford: Oxford University Press.

Gilquin, G. (2006). The place of prototypicality in corpus linguistics: Causation in the hot seat. In S. Gries & A. Stefanowitsch (Eds.), Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis. Berlin: Mouton de Gruyter, 159-191.

Gilquin, G. (2008). What you think ain't what you get: Highly polysemous verbs in mind and language. In J.-R. Lapaire, G. Desagulier & J.-B. Guignard (Eds.), From gram to mind: Grammar as cognition. Bordeaux: Presse Universitaires de Bordeaux, 235-255.

Glynn, D. (2014). Polysemy and synonymy: Cognitive theory and corpus method. In D. Glynn & J. A. Robinson (Eds.), Corpus methods for semantics: Quantitative studies in polysemy and synonymy. Amsterdam: Benjamins, 7-38.

Gries, S. Th. (2006). Corpus-based methods and cognitive semantics: The many senses of to run. In S. Gries & A. Stefanowitsch (Eds.), Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis. Berlin: Mouton de Gruyter, 57-99.

Heylen, K., Tummers, J. & Geeraerts, D. (2008). Methodological issues in corpus-based Cognitive Linguistics. In Kristiansen & R. Dirven (Eds.), Cognitive Sociolinguistics: Language variation, cultural models, social systems. Berlin: Mouton de Gruyter, 91-128.

Lindqusit, H. (2009). Corpus linguistics and the description of English. Edinburgh: Edinburgh University Press.

McEnery, T. & Gabrielatos, C.. (2006). English corpus linguistics. In B. Aarts & A. McMahon (Eds.), The handbook of English linguistics. Malden, Massachusetts: Blackwell, 33-71.

McEnery, T. & Wilson, A.. (2001). Corpus linguistics. 2nd edn. Edinburgh: Edinburgh University Press.
Nordquist, D. (2004). Comparing elicited data and corpora. In M. Achard and S. Kemmer (Eds.), Language, culture and mind. Stanford: CSLI Publications, 211-24.
Rosch, E. (1973). Natural categories. Cognitive Psychology, 4 (3), 328-50.
Rosch, E. (1975a). Cognitive reference points. Cognitive Psychology, 7, 532-47.
Rosch, E. (1975b). Cognitive representations of semantic categories. Journal of Experimental Psychology, 104 (3), 192-233.
Schmid, H.-J. (2007). Entrenchment, salience and basic levels. In D. Geeraerts & H. Cuyckens (Eds.), The Oxford Handbook of Cognitive Linguistics. Oxford: Oxford University Press, 117-38.
Taylor, J. (2003). Linguistic Categorization. 3rd edn. Oxford: Oxford University Press.
Taylor, J. R. (2012). The mental corpus: How language is represented in the mind. Oxford: Oxford University Press.
Werner, J. & Mukherjee, J. (2012). Highly polysemous verbs in New Englishes: A corpus-based pilot study of Sri Lankan and Indian English. In S. Hoffman (Ed.), English corpus linguistics: Looking back, moving forward. Amsterdam: Rodopi, 249-66.

• • •

Ngum Meyuhnsi Njende
University of Leuven
ngummeyuhnsi.njende@kuleuven.be

Kristin Davidse
University of Leuven
kristin.davidse@kuleuven.be

Enumerating *there*-clauses and *there*-clefts: a corpus-based study

This paper wants to clarify the controversial status of two related English constructions: enumerative *there*-clauses, e.g. (1), whose distinctness from ordinary existential clauses like (3-4) has been disputed, and enumerative *there*-clefts (2), which are not generally recognized as specificational clefts.

(1) His nomination of formative influences gives a clue. There is Brian Ashton and Jack Rowell. (WB Times)
(2) Now you ,ve got a fair sort of permanent staff now. There's Herman has been there for years. (WB BrSpoken)

An argument often adduced for the difference between enumerative and ordinary existentials is the typical definiteness of the NPs in the former, e.g. *Brian Ashton and Jack Rowell* (1) and the 'indefiniteness' restriction said to apply to the latter, e.g. *some nasty girls* in (3) (Lyons 1977). As shown by non-enumerative examples like (4), where *the usual drunk Scotsman* evokes 'a' (new) instance of the familiar type *drunk Scotsman*, this restriction is of a semantic-pragmatic nature. Ward & Birner (1995) correctly point out that examples like (3) and (4) as well as listing examples like (1) all share the pragmatics of presenting entities that are in some sense 'new'. However, for this reason they reject the distinction between enumerative an non-enumerative existentials. Against this, we argue that the different coded meanings of the two types should be recognized. Ordinary existentials like (3-4) introduce unidentifiable instances of the type designated by the common noun and modifiers in the Existent NP, *nasty girls* (3) and *drunk Scotsman* (4) respectively. By contrast, enumerative existentials like (1) name one or more – typically identifiable – instances of a superordinate type to be retrieved from the preceding text: Ashton and Rowell are instances of *formative influences*.

(3) I went to an all-girls' school and there were (*the) some nasty girls there. (WB)
(4) There was the usual drunk Scotsman in the corner. (WB)

A further argument for recognizing enumerative existentials as a distinct type is the features they share with enumerative *there*-clefts like (2), whose focal NP likewise names one or more – typically identifiable – instances, e.g. *Herman* in (2), of a superordinate type, which is retrievable from the preceding text, permanent staff, but which is also explicitly coded by the cleft relative clause, *has been there for years*. This tallies with Lambrecht's (2001) analysis of examples like (2) as specificational clefts whose focal NP non-exhaustively lists a value for the variable 'x' in the open proposition contained in the cleft relative clause – an analysis not subscribed to by Huddleston & Pullum (2002:1396).

We will verify and further develop the posited similarities in a corpus-based description, filtering extractions of the query [tag="EX"][lemma="be"][tag="NP|PP|DET"] from BrSpoken and Times in WordbanksOnline (WB). For enumerative *there*-clauses and *there*-clefts, we will analyse and quantify the occurrence of (i) simple and coordinated NPs, and (ii) referring expressions: proper names, pronouns, definite or indefinite NPs, and functionally interpret these in function of the listing semantics. For *there*-clefts, we will assess whether the proposition in the relative clause is textually evoked, inferrable, anchored or brand-new in relation to the preceding text (Kaltenböck 2004, Gentens 2016). We will then extend this analysis to enumerative *there*-clauses: it should always be possible to make the superordinate type explicit in an added relative clause, e.g. (1) *There is Brian Ashton and Jack Rowell that are formative influences*, which will allow us to assess whether the superordinate type is textually evoked, inferable, or possibly anchored in the preceding discourse. Presumably it is never brand-new, as it would then be impossible to decide which superordinate type the enumerative existential lists instances of. This will give us an insight into the

discursive motivations for choosing an enumerative *there*-clause or an enumerative *there*-cleft.

References

Gentens, C. (2016). The discursive status of extraposed object clauses. Journal of Pragmatics 96, 15-31.

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

Kaltenböck, G. (2004). It-extraposition and Non-extraposition in English: a Study of Syntax in Spoken and Written Texts. Wien: Braumüller.

Lambrecht, K. (2001). A framework for the analysis of cleft constructions. Linguistics 39, 463-516.

Lyons, J. (1977). Semantics. Cambridge: Cambridge University Press.

Ward, G. & Birner, B. (1995). Definiteness and the English existential. Language 71, 722-742.

• • •

Paloma Núñez-Pertejo
University of Santiago de Compostela
pnunez.pertejo@usc.es

Practically impossible: From 'Practicality' to 'Approximation'

Considerable attention has been paid in the literature to *–ly* adverbs from a variety of different perspectives, both synchronic and diachronic (e.g. Aijmer 2011; Defour 2012). Yet, the degree adjunct *practically*, a 'downtoner' (cf. Quirk et at. 1985: 445) of the so-called 'approximating' sub-group, like *almost* or *virtually* (cf. Huddleston & Pullum 2002: 720-724; cf. also Quirk *et al.* 1985: 597-602), has not been the object of much discussion. *Practically* has undergone a line of development similar to most *–ly* adverbs, and hence emerges as a manner adverb with the meaning 'in a practical manner', 'in practice' (frequently opposed to 'theoretically' or 'speculatively'), then evolving into a degree modifier of the approximating type with the meaning 'almost', 'in effect', 'nearly'. Following this process of subjectification (cf. Swan 1997; Traugott 1995), this paper seeks to contribute to the history of *practically* by using a corpus-based methodology. To this end, the *Corpus of Late Modern English Texts, version 3.0* (CLMET3.0; 34 million words; cf. De Smet, Diller & Tyrkkö 2011) will be used, supplemented with additional evidence from the *Old Bailey Corpus*, *version 2.0* (OBC2.0; 24 million words).

According to the OED, *practically* was first attested as a manner adverb in 1571 (*OED* s.v. *Practically*, adv. 1), and as a degree modifier almost two centuries later (*OED*

s.v. *Practically*, adv. 2). Nevertheless, a preliminary analysis shows that it was a rather infrequent adverb in British English as late as the beginning of the 18th century, with only 11 tokens recorded in the first sub-period of CLMET3.0 (1710-1780), while the first instances in the OBC2.0 do not appear until the end of the 19th century. Further findings seem to point to the infrequency of *practically* as a manner adverb in contemporary British English, whereas its role as a degree modifier, on the other hand, seems to be predominant. However, in contrast to other *–ly* adverbs such as *absolutely* or *totally*, *practically* has not moved further along in the cline of directionality proposed by Traugott (1982; cf. also Traugott & Dasher 2002) in that it very rarely functions as an independent response token, which may have to do with the nature of the sub-group of degree modifiers to which *practically* belongs, i.e. 'approximating', as opposed to adverbs like *absolutely* or *totally*, which belong to the 'maximal' sub-group usually referred to as 'maximizers'.

In addition to the analysis of the evolution of *practically* from manner adverb to degree modifier, and given that this adverb seems to trigger "a strong negative implicature" (Huddleston & Pullum 2002: 723), I will explore the range of collocational patterns that *practically* allows, as well as the various syntactic functions it may have in the sentence.

References

Aijmer, K. (2011). Are you totally spy? A new intensifier in Present-day American English. In S. Hancil (Ed.), Marqueurs discursifs et subjectivité. Publications des Universités de Rouen et de Havre, 155-172.

Defour, T. (2012). The pragmaticalization and intensification of verily, truly and really. A corpus-based study on the developments of three truth-identifying adverbs. In M. Markus, Y. Iyeiri, R. Heuberger & E. Chamson (Eds.), Middle and Modern English Corpus Linguistics: A Multidimensional Approach. Amsterdam/Philadelphia: John Benjamins, 75-92.

De Smet, H., Diller, H. J. & Tyrkkö, J. (2011). The Corpus of Late Modern English Texts, version 3.0 (CLMET3.0). More information: https://perswww.kuleuven.be/~u0044428/clmet3_0.htm

Huber, M., Nissel, M. & Puga, K. (2016). Old Bailey Corpus 2.0. hdl:11858/00-246C-0000-0023-8CFB-2

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

OED (Oxford English Dictionary). 3rd edn. in progress: OED Online, March 2000-, ed. John A. Simpson. <www.oed.com>

Quirk, R., Greenbaum, S., Leech, G. N. & Svartvik, J. (1985). A Comprehensive Grammar of the English Language. London/New York: Longman.

Swan, T. (1997). From Manner to Subject Modification: Adverbialization in English. Nordic Journal of Linguistics 20, 179-195.

Traugott, E. C. (1982). From Propositional to Textual and Expressive Meanings; Some Semantic-Pragmatic Aspects of Grammaticalization. In W. P. Lehmann

& Y. Malkiel (Eds.), Perspectives on Historical Linguistics. Amsterdam: John Benjamins, 245-271.
Traugott, E. C. (1995). Subjectification in Grammaticalisation. In D. Stein & S. Wright (Eds.), Subjectivity and Subjectification: Linguistic Perspectives. Cambridge: Cambridge University Press, 31-54.
Traugott, E. C. & Dasher, R. (2002). Regularity in Semantic Change. Cambridge: Cambridge University Press.

• • •

Ignacio M. Palacios Martínez
University of Santiago de Compostela
ignacio.palacios@usc.es

***Help me move to that, blood*. A corpus-based study of the syntax and pragmatics of vocatives in the language of British teenagers**

Braun's monograph (1988) on different terms of direct address, or vocatives, across several languages prepared the ground for the emergence of a large body of work dealing with the form, sentence position and pragmatic function of these lexical items in English. Some of these studies (Leech 1999, Murphy & Farr 2012) have focused on the distribution and function of vocatives in different varieties of English, while others have addressed a single form (Kiesling 2004, Rendle-Short 2010), or their behaviour in particular text types, such as telephone calls in radio programmes (McCarthy & O'Keefe 2003), family discourse (Clancy 2015) or political debates (Jaworski & Galasiński 2000).The literature also includes a number of contrastive studies (Axelson 2007, Hill 2007, Alba-Juez 2009, Heyd 2014, Stenström 2015).

In this paper I will consider one specific group of English vocatives, the so-called "familiarisers" (Leech 1999, Biber et al. 1999), which include examples such as the following:

(1) every day innit *boy* no *man* just look *bruv*. (LEC)
(2) that's your little brother *bruv* innit. (LEC)
(3) just using phrases like 'safe' and 'yes blood' and all that he just speaks like a black. (LEC)
(4) he goes 'nah *mate*'. (COLT)

Adopting a corpus-based methodology, one which allows for the analysis and comparison of data from several corpora (the *Bergen Corpus of London Teenage English*, the *London English Corpus* and the *British National Corpus*), I will explore:

(a) The extent to which these forms are typical of the language of youths and, on the contrary, are not so frequent in adult speech;
(b) The different lexical items that can be included in this general category of familiarisers;
(c) Their position in the sentence and the significance of this position in terms of pragmatic meaning;
(d) Their semantics and main pragmatic functions;
(e) Any notable changes in the evolution of these vocatives over time; and
(f) Possible reasons explaining their high occurrence in teen talk.

Preliminary findings indicate that these forms are more common in the expression of British teenagers than in that of adult speakers, by a factor of more than six in the present data. Some of the most frequent are: *man*, *bru/bro*, *mate* and *boy*. The first two will receive the most attention in this paper. As expected, the majority of these items occur in final position. Pragmatically speaking, they can carry a wide variety of functions: from summoning attention to creating solidarity and comradeship among the speakers, the latter being quite frequent in teen talk. We even find more than one of these vocatives together in the same turn (1) or used with a pragmatic marker (2), and some are also used metalinguistically (3); they are also prevalent in reported speech (4). A few of these familiarisers, such as *bruv/bro*, are found in the most recent corpora (*London English Corpus*) while others, like *bloke*, *pal*, f*olk*, *sister* and *woman*, occur only in COLT and the BNC. Finally, in most cases these vocatives have lost their deictic reference to the addressee and have in fact become pragmatic markers, or may even adopt features of another category, such as pronouns (Cheshire 2013).

References

Alba-Juez, L. 2009. Little words in Small Talk: Some considerations on the use of the pragmatic markers man in English and macho/tío in Peninsular Spanish. In R. Leow, H. Campos & D. Lardiere (Eds), Little Words. Their History, Phonology, Syntax, Semantics, Pragmatics, and Acquisition, 171-181.

Axelson, E. (2007). Vocatives: A double-edged strategy in intercultural discourse among graduate students. Pragmatics 17(1), 95-122.

Biber, D., Stig J., Geoffrey L., Susan C. & Finegan, E. 1999. Longman Grammar of Spoken and Written English. Harlow: Pearson.

Braun, F. (1988). Terms of Address: Problems of Patterns and Usage in Various Languages and Cultures. Berlin: Mouton de Gruyter.

Cheshire, J. (2013). Grammaticalisation in social context: The emergence of a new English pronoun. Journal of Sociolinguistics 17(5), 608-633.

Clancy, B. (2015). "Hurry up baby son all the boys is finished their breakfast". Examining the use of vocatives as pragmatic markers in Irish Traveller and settled family discourse. In C. P. Amador-Moreno, K. McCafferty & E. Vaughan (Eds), Pragmatic Markers in Irish English. Amsterdam: John Benjamins, 229-247.

Heyd, Th. (2014). Dude, Alter! A tale of two vocatives. Pragmatics and Society 5(2), 271-295.

Hill, V. (2007). Vocatives and the pragmatic-syntax interface. Lingua 117, 2077-2105.
Kiesling, S. (2004). Dude. American Speech 79(3), 281-305.
Leech, G. (1999). The distribution and function of vocatives in American and British English conversation. In Hilde Hasselgärd & Signe Oksefjell (Eds), Out of Corpora. Studies in Honour of Stig Johansson. Amsterdam: Rodopi, 107-120.
McCarthy, M. & O'Keeffe, A. (2004). What's in a name? Vocatives in casual conversations and phone-in-calls. In P. Leistyna & C. Meyer (Eds), Corpus Analysis: Language Structure and Language Use. Amsterdam: Rodopi, 153-185.
Murphy, B. & Farr, F. (2012). The use of vocatives in spoken Irish English. In B. Migge & M. N. Chiosáin (Eds), New Perspectives on Irish English. Amsterdam: John Benjamins, 203-224..
Rendle Short, J. (2010). Mate as a term of address in ordinary interaction. Journal of Pragmatics 42(5): 1201-1218.
Stenström, A.-B. (2015). Teenage Talk: From General Characteristics to the Use of Pragmatic Markers in a Contrastive Perspective. London: Palgrave Macmillan.

• • •

Magali Paquot
FNRS - Université catholique de Louvain
magali.paquot@uclouvain.be

Hubert Naets
Université catholique de Louvain
hubert.naets@uclouvain.be

The role of the reference corpus in studies of EFL learners' use of statistical collocations

In learner corpus research (LCR), there has been a recent boom in the number of studies that have investigated English as a Foreign Language (EFL) learners' use of statistical collocations (e.g. Bestgen & Granger 2014; Granger & Bestgen 2014; Paquot & Naets, 2015; Paquot, forthcoming a & b). These studies have adopted an approach first put forward by Schmitt and colleagues (e.g. Durrant & Schmitt 2009) to assess whether and to what extent the word combinations used by learners are 'native-like' by assigning to each pair of words in a learner text an association score (typically a pointwise mutual information and/or a t-score) computed on the basis of a large reference corpus.

The reference corpus differs across studies. Thus, Granger & Bestgen (2014) made use of the British National Corpus (BNC) to evaluate EFL learners' use of bigrams in the International Corpus of Learner English (Granger et al 2009); Paquot (forthcom-

ing a & b) extracted statistical collocations from the L2 Research Corpus (L2RC), i.e. a large specialized corpus of research articles in applied linguistics, to assess learners' use of adjective + noun and verb + object combinations in term papers in linguistics written by French EFL learners sampled from the Varieties of English for Specific Purposes Database (VESPA); and Paquot & Naets (2015) used the web corpus ENCOW14 (http://corporafromtheweb.org/encow14/) to analyze statistical collocations in the Longitudinal Database of Learner English (LONGDALE, Meunier, 2015).

The main objective of this study is to investigate the role of the reference corpus in LCR studies of statistical collocations in learner writing. It is driven by the following research questions:

- To what extent are results replicable if another reference corpus is used to calculate association scores?
- Depending on the learner corpus data investigated, should we use a general reference corpus or a specialized corpus to compute association measures?

To answer our research questions, we replicate the method used in Paquot & Naets (2015) and Paquot (forthcoming a & b): we extract relational co-occurrences (i.e. adjective + noun, adverb + adjective, adverb + verb and verb + direct object relations) from dependency parsed versions of the BNC, ENCOW14 and L2RC and compute their mutual information (MI) scores with the Ngram Statistics Package (NSP). We then use MI scores computed on the basis of the three reference corpora to analyze the same relational co-occurrences in learner texts rated at different CEFR levels (i.e. B2, C1, C2) sampled from ICLE and VESPA. We compute mean MI scores for each dependency relation in each learner text (cf. Bestgen & Granger 2014) and compare their distribution across proficiency levels. Distributions in the CEFR-based learner data sets are tested for normality and accordingly compared with ANOVAs followed by Tuckey contrasts or Kruskal-Wallis rank sum tests followed by pairwise comparisons using Wilcoxon rank sum tests.

Preliminary results confirm previous research by demonstrating that the more advanced learners use more native-like collocations irrespective of the reference corpus. However, MI scores computed on the basis of the three different reference corpora seem to reveal different aspects of phraseological proficiency in learner writing, most notably the use of general vs. genre-specific collocations.

References

Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. Journal of Second Language Writing, 26, 28–41.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? IRAL - International Review of Applied Linguistics in Language Teaching, 47(2), 157–177. doi:10.1515/iral.2009.007

Granger, S. & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. International Review of Applied Linguistics in Language Teaching, 52(3): 229-252.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (2009). The International Corpus of Learner English. Version 2. Handbook & CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain, 2009.
Meunier, F. (2015) Introduction to the LONGDALE project. In E. Castello, K. Ackerley, & F. Coccetta (Eds.) Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment. Bern: Peter Lang.
Paquot, M. (forthcoming a). Phraseological competence: a missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. Language Assessment Quarterly.
Paquot, M. (forthcoming b). The phraseological dimension in interlanguage complexity research. Second Language Research.
Paquot, M., Hasselgård, H. & Oksefjell Ebeling, S. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds) Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain.
Paquot, M. & Naets, H. (2015). Adopting a relational model of co-occurrences to trace phraseological development. Paper presented at the 3rd Learner Corpus Research Conference, 11-13 September 2015, The Netherlands.

• • •

Laura Paterson
The Open University
laura.paterson@open.ac.uk

Visualising corpora using Geographical Text Analysis: (Un)employment in the UK, a case study

Geographical Text Analysis (GTA) is way of analysing language which focuses on the geographical locations referred to in texts. GTA centres on corpus analysis and involves the manipulation of concordance lines. Once a corpus has been compiled and query terms have been selected, concordance lines are generated and are fed through a software programme known as a geoparser.

Geoparsing involves the automatic detection of place names occurring within a set span L/R of the concordance node. Once identified these Place Name Collocates (PNCs) are tagged with the geographical coordinates corresponding to that place's location on the Earth's surface. The geoparsed output is uploaded into a Geographical Information System (GIS) – cartographic database software – to facilitate the creation of maps. Thus, GTA allows the researcher to visualise their corpus cartographically

and adds a consideration of physical space to language analysis. Furthermore, these maps, and the data which underpins them, can be compared to other quantitative (and visualisable) sources, such as official statistics and census data.

• • •

John Payne
University of Manchster
john.payne@manchester.ac.uk

Eva Berlage
Universität Hamburg
Eva.Berlage@uni-hamburg.de

Semantic relations in the genitive alternation

The genitive alternation is, without doubt, one of the most well-researched cases of grammatical variation to date (for an overview see Rosenbach 2002, 2005, 2014; Szmrecsanyi et al. 2013). Most studies converge on the conclusion that animacy of the dependent contributes most to the explained variation, with weight and discourse status also significant (O'Connor et al. 2013; Börjars et al. 2013).

A small number of studies have argued, however, that the semantic relation that holds between head (*car*) and dependent (*man*) in an example such as *the man's car* is one of the major determinants governing the choice of variants (see e.g. Gries 2002; Stefanowitsch 2003; Szmrecsanyi 2013). In studies in which the semantic relations involved in the genitive alternation are taken into account, they are generally divided into two larger groups based either on the opposition alienable vs inalienable (see e.g. Nichols 1986:77; 1988: 572-573), or the notion prototypical vs non-prototypical, with the ownership relation being the most prototypical (see e.g. Taylor 1989: 679–84, 2003: 228–31). However, at a finer level of description, there are potentially about twenty distinct semantic relations for which variation seems to be permitted (see Payne et al. 2013). An indication that it might be useful to incorporate this finer level of distinctions into a multivariate analysis is provided by Payne & Berlage (2011), who argue that the effect of semantic relations on the genitive variation is so strong that it results in varying preferences for the *s*-genitive even amongst those semantic relations which have hitherto been considered prototypical. These relations include ownership (*the man's car*), bodypart (*the girl's eyes*) and kinship (*the child's parents*).

In this paper, using data from the BNC and adding to the semantic relations used by Payne & Berlage (2011), we investigate further the extent to which individual semantic relations contribute to predicting the genitive alternation. While Payne &

Berlage (2011) is restricted to five different semantic relations (ownership, bodypart, kinship, interpersonal relation and depiction), this paper adds the following four relations: undergoer (e.g. *the interpretation of the book*), property (e.g. *the girl's beauty*), agent (e.g. *the man's arrival*) and creator (e.g. *the manager's agenda*). The hypothesis explored is that, rather than there being a coarse distinction between prototypical and non-prototypical relations, there exists a finer-grained hierarchy of semantic relations where ownership is indeed at the top of the list, while depiction (e.g. *photo of the man* where the man is on the photo) is at the bottom (strongly favouring the *of*-PP construction while at the same time not excluding the *s*-genitive variant).

We conclude by proposing a uniform account of the hierarchy, namely that those semantic relations at the top of the list relate to referents that we have most control over (our possessions), whilst those at the bottom involve no element of control at all. This explanation in terms of decreasing degrees of 'control' applies not only to cases that allow for variation between the *s*-genitive and the *of*-PP construction, but also extends to those that invariably take the latter construction (e.g. *glass of water*).

References

Börjars, K., Denison, D., Krajewski, G. & Scott, A. (2013). Expression of possession in English: the significance of the right edge. In Börjars, K., D. Denison & A. Scott (eds.), Morphosyntactic Categories and the Expression of Possession. Amsterdam: Benjamins, 123-148.

O'Connor, C., Maling, J. & Skarabela, B. (2013). Nominal categories and the expression of possession: a cross-linguistic study of probabilistic tendencies and categorical constraints. In K. Börjars, D. Denison & A. Scott (eds.), Morphosyntactic Categories and the Expression of Possession. Amsterdam: Benjamins, 89-122.

Gries, S. Th. (2002). Evidence in linguistics: three approaches to genitives in English. In R. M. Brend, W. J. Sullivan & A. R. Lommel (eds), Lacus Forum XXVIII: What Constitutes Evidence in Linguistics? Fullerton, CA: LACUS, 17-31.

Nichols, J. (1986). Head-marking and dependent-marking grammar. Language 62(1), 56–119.

Nichols, J. (1988). On alienable and inalienable possession. In W. Shipley (ed.), In Honor of Mary Haas. From the Haas Festival Conference on Native American linguistics. Berlin/New York: Mouton, 557-609.

Payne, J. & Berlage, E. (2011). The effect of semantic relations on the genitive variation. Paper presented at ISLE 2, University of Boston, 17-21 June 2011.

Payne, J., Pullum, G. K., Scholz, B. & Berlage, E. (2013). Anaphoric one and its implications. Language 89, 794–829.

Rosenbach, A. (2002). Genitive Variation in English. Conceptual Factors in Synchronic and Diachronic Studies. Berlin/New York: Mouton de Gruyter.

Rosenbach, A. (2005). Animacy versus weight as determinants of grammatical variation in English. Language 81(3), 143–89.

Rosenbach, A. (2014). English genitive variation: the state of the art. In J. Payne & E. Berlage (eds.) 2014. Special Issue on the Genitive Variation in English. English Language and Linguistics 18.2, 215–262.

Stefanowitsch, A. (2003). Constructional semantics as a limit to grammatical alternation. In G. Rohdenburg & B. Mondorf (eds.), Determinants of Grammatical Variation in English. Berlin/New York: Mouton de Gruyter, 413-443.
Szmrecsanyi, B. (2013). The great regression: Genitive variability in Late Modern English news texts. In K. Börjars, D. Denison & A. Scott (eds.), Morphosyntactic Categories and the Expression of Possession. Amsterdam: Benjamins, 59-88.
Taylor, J. (1989). Possessive genitives in English. Linguistics 27, 663–86.
Taylor, J. (2003). Linguistic Categorization. Third edition. Oxford: Oxford University Press.

• • •

Florent Perek
University of Birmingham
f.b.perek@bham.ac.uk

Periodization of constructional productivity in diachronic corpora

Studies of grammatical change typically describe diachronic developments in terms of discrete stages: periods of relative stability and times of shift in the recorded usage of some grammatical construction. In recent work, Gries & Hilpert (2008) proposed a quantitative usage-based approach to this issue, called variability-based neighbour clustering (VNC). VNC is a customized version of agglomerative clustering that consists in aggregating adjacent periods that are closely similar in terms of some quantitative criteria measured on occurrences of a construction in diachronic corpora. The output of VNC is a partition of the time scale into periods that are maximally coherent with respect to the relevant criteria.

Most applications of VNC so far have taken as their basis purely quantitative information: token frequency, type frequency and other measures derived from them, or the frequency distribution of lexemes occurring in a slot of a construction. However, information of this kind does not directly capture qualitative stages of productivity, i.e., the different semantic classes of lexical items used in a construction over time.

This paper presents an extension of VNC that addresses these limitations by drawing on a distributional semantic model as a proxy to word meaning. Drawing on the observation that words occurring in similar contexts tend to have similar meanings, distributional semantic representations approximate the meaning of a word by recording its co-occurrence with other words in a vast corpus (Turney & Pantel 2010). The present approach consists in adding the distributional representations of all words occurring in a construction at different points in time, and using the resulting combinations as input to VNC.

The method is illustrated by two case studies. The first case study on verbs used in the "Verb *the hell out of* NP" construction (e.g. *You scared the hell out of me*) shows that the semantic development of a construction does not always match that of its quantitative aspects, like token or type frequency. The second case study on the verbs used in the *way*-construction (e.g. *They pushed their way through the crowd*) compares the results of the present method with those of collostructional analysis (Hilpert, 2006). While the results overlap to some degree, in that both methods identify successive periods of productivity during which the construction has gradually attracted more abstract classes of verbs, it is shown that the present approach measures semantic change with greater precision, both regarding the nature of changes and their chronology. In sum, this method offers a promising exploratory approach to capture variation in the semantic range of lexical fillers of constructions and model constructional change.

References

Gries, S. & Hilpert, M. (2008). The Identification of Stages in Diachronic Data: Variability-based Neighbor Clustering. Corpora, 3, 59–81.

Hilpert, M. (2006). Distinctive collexeme analysis and diachrony. Corpus Linguistics and Linguistic Theory, 2(2). 243–57.

Turney, P. & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37, 141–188.

• • •

Maura Ratia
University of Helsinki
maura.ratia@helsinki.fi

Carla Suhr
University of Helsinki
carla.suhr@helsinki.fi

Interpreting the world of late modern English medical writing: From *I witnessed* to *it was observed*?

In the history of English, the register of medical writing has been established as the spearhead in changing scientific thought-styles (see eg. Taavitsainen and Pahta 1995). Research into this register has shown a shift from logocentric, authority-driven discourse in the late Middle Ages to a more empirically motivated discourse that elevates personal experience over classical authorities in the early modern period (Pahta and Taavitsainen 2011). However, medical writing between 1700 and the late 20th century

has only recently begun to be studied in depth, mainly because of a lack of corpus material suitable for linguistic study. This paper begins to tackle the gap by focusing on medical writing in the 18th century – the century that saw, for example, the first academic journals dedicated to medicine. This paper investigates the use of first-person pronouns and passive constructions in LMEMT, the Corpus of Late Modern English Medical Texts (Taavitsainen et al. forthcoming), a collection of writings in the medical register totaling over 2 million words. The LMEMT corpus provides a tool for analytical assessment of how changes proceed with different paces in different layers of writing, such as in texts written for professional or for lay audiences. For a diachronic perspective, the same searches are carried out in EMEMT, the Corpus of Early Modern English Medical Texts (Taavitsainen et al 2010), and in the Medicor corpus, which comprises of medical texts from the 1980s and 1990s (see Vihla 1999).

The emphasis on empirical, strictly controlled experiments as sources of medical knowledge that developed in the early modern period continues today, but the style of writing has changed: present-day medical writing is very abstract and impersonal, and first-person pronouns are rare (Vihla 1999). By focusing on first-person pronouns, on the one hand, and passive constructions, on the other hand, in these three corpora we will be able to see whether the shift towards modern style begins already in the 18th century. The contexts of use of these features are determined to identify patterns of use, which makes it possible to determine what kinds of medical texts lead the change and how it spreads into various layers of writing dependent on, for instance, audience or genre. At the same time, we will be able to show where conventions already established earlier continue in the late modern period and perhaps even into the present day.

References

EMEMT: Early Modern English Medical Texts. 2010. Compiled by I. Taavitsainen, P. Pahta, M. Mäkinen, T. Hiltunen, V. Marttila, M. Ratia, C. Suhr & J. Tyrkkö. Amsterdam: John Benjamins.

LMEMT: Late Modern English Medical Texts. Forthcoming. Compiled by I. Taavitsainen, T. Hiltunen, A. Lehto, V. Marttila, R. Oinonen, P. Pahta, M. Ratia, C. Suhr & J. Tyrkkö.

Pahta, P. & Taavitsainen, I. (2011). An interdisciplinary approach to medical writing in Early Modern English. In I. Taavitsainen & P. Pahta (eds.), Medical Writing in Early Modern English, , 9–29. Cambridge: Cambridge University Press.

Taavitsainen, I. & Pahta, P. (1995). "Scientific 'thought-styles' in discourse structure: Changing patterns in a historical perspective." In B. Warvik, S. K. Tanskanen & R. Hiltunen (eds.), Organization in discourse: Proceedings from the Turku Conference, 519–29. University of Turku.

Vihla, M. (1999). Medical Writing. Modality in Focus. Amsterdam and Atlanta: Rodopi.

• • •

Antoinette Renouf
Birmingham City University
antoinette.renouf@bcu.ac.uk

The life-cycle of a term in mainstream UK news text

We have previously modelled the life-cycle of a word in text (Renouf 2013), identifying the stages of activity and change which it undergoes in a large corpus of UK news text across time. Our objective in this paper is to observe the evolution of a term from a similar perspective. A term is a special case, a word or phrase which names a concept or object within a specialised domain of study or activity. It is a semantically fixed entity at the time of coinage (Sager 1994). Terms vary in kind, from formulae such as *H2N2*, to classical formations such as *infectious mononucleosis*, to scientific discoveries like *event horizon*, to 'lay' medical terms like *Asian flu*, to terms closer to general language, such as *mouse* or *desktop* in computing. Within a mainstream newspaper, a term occurs in two main locations: in the publication's sections of specialised text; or in the general news section if, like *HS2*, it reflects a real world event or, like *PVC*, has established its place there over time.

The approach is based on the theory that the life-cycle of a term in diachronic news text will actually differ from that of a word, for a number of reasons. Firstly, terms come into being somewhat differently. A word is typically created by general word formation rules, whereas a term is typically based on a formula, classical roots or linguistic components within a specialised domain, or it undergoes "terminologisation", combining pre-existing general words to create a specialised meaning, such as *standalone* or *downsize*. Secondly, a word typically resides in general text, whereas a term may take one of several paths. If highly specialised, it will probably remain in its rarefied domain, quoted occasionally in general text. If it is less rarefied, a layman's term, it may move into general use, where its underlying concept will probably acquire a "conceptual fuzziness" (Halskov 2005). Alternatively, a popular term may undergo a process of "determinologisation" (Meyer & Mackintosh, 2000) in everyday use, gaining a new metonymic or metaphorical sense, as with *water-cooler*, in *water-cooler conversation*.

In the paper, we shall study a selection of terms, partly informed by the work of Condamines & Picton (2014) and Meyer & Mackintosh (ibid.), and explore their paths through the corpus, observing features of terminological and determinological evolution, including the co-existence of evolving meanings and uses, 'primary and secondary' determinologisation, inflectional specificity versus lemmatisation, the creative use of layman's and specialist terms, and the issue of grammatical change. The description of these features forms the structure of the paper's findings. The corpus in question is a 1.5 billion-word diachronic collection of Independent and Guardian texts, from 1984-2015, processed by the WebCorp Linguist's Search Engine (http://wse1.webcorp.org.uk/). It is hoped that this corpus-based, diachronic perspective will help to

clarify the evolution of terms for terminologists and translators, and throw new light on the nature of semantic change in terminology for corpus linguists generally.

References

Condamines, A. & Picton, A.. (2014). Étude du fonctionnement des nominalisations déverbales dans un contexte de déspécialisation. In SHS Web of Conferences, July 2014, Vol.8, 697-71.

Halskov, J. (2005). Probing the Properties of Determinologization - the DiaSketch. In 2008-2016 researchgate.net.2.

Meyer, I. & Mackintosh, K. (2000). When terms move into our everyday lives: An overview of de-terminologization. Terminology, 6(1), 111 –138.

Renouf, A. (2013). A Finer Definition of Neology in English: the life-cycle of a word. In H. Hasselgård, S. Oksefjell Ebeling & J. Ebeling (Eds.), Corpus Perspectives on Patterns of Lexis. Amsterdam: John Benjamins, 177-208.

Sager, J. C. 1994. Terminology: Custodian of Knowledge and Means of Knowledge Transfer. Terminology 1(1), 7-15.

• • •

Patricia Ronan
TU Dortmund
patricia.ronan@udo.edu

Gerold Schneider
University of Zürich
gschneid@ifi.uzh.ch

Directive Speech Acts in SPICE Ireland and beyond

This paper offers an investigation of directive speech acts in SPICE Ireland. On the basis of Kallen and Kirk (2012), the study asks which qualitative and quantitative subdivisions are found in directive speech acts in SPICE Ireland and how these may compare to corpus data from other varieties of English, especially British English on the basis of ICE-GB (Nelson, G., S. Wallis, and B. Aarts 2002).

Our aim is to offer a sub-classification of directives in SPICE Ireland in order to determine how direct directives compare qualitatively and quantitatively to indirect directives, and to other directives. Irish English has repeatedly been classified as a variety that is particularly given to avoiding face-threatening acts (Barron 2008, Kallen 2005), therefore a preference of indirect over direct directives could be expected. The paper further aims to compare the results to data from other ICE corpora.

Unfortunately no pragmatically annotated corpora exist for other ICE varieties, therefore we approximate by using a machine learning approach.

Thus the proposed paper will first extract directives from relevant subsections of the pragmatically annotated SPICE Ireland corpus (Kallen and Kirk 2012). The focus will particularly be on unscripted interpersonal communication, such as classroom discussions, face-to face conversations, broadcast discussions and telephone conversations, and collected examples will be evaluated manually. The extracted data will be classified into different categories as discussed by Kallen and Kirk (2012: 30), who find directives in the form of direct orders, partly with speech act verbs such as *order* or *command*, as well as indirect directives in questions and tag questions. Further instances will be found in contextually determined directives such as

1. So that's the sixteenth of August then that Tuesday <P1B-002$B> (Kallen and Kirk, ibid.)

which stems from classroom discussions and, even though it is not formally marked as an order, the content determines a hand-in date for student essays and thus constitutes a directive (Kallen and Kirk, ibid.).

Second, a machine learning approach is used to determine the approximate counts of directives in other varieties of ICE, and we then compare the results to the data from SPICE Ireland. For this, we apply a bag-of-word-model and use state-of-the-art algorithms such as logistic regression, and evaluate its performance. This allows us to also offer a preliminary qualitative and quantitative comparison of directives in the two selected ICE corpora. In this, the paper also answers Barron's (2008: 58) call to compare directness on the level of speech acts.

References

Barron, A. (2008). The Structure of Requests in Irish English and English English. In: A. Barron & K. P. Schneider (Eds.), Variational Pragmatics: A focus on regional varieties in pluricentric languages, 35-67.

Kallen, J. (2005). Politeness in Ireland: '…In Ireland, It's Done With- out Being Said'. In: L. Hickey & M. Stewart (Eds.) Politeness in Europe. Clevedon: Multilingual Matters, 130-144.

Kallen, J. & Kirk, J. (2012). SPICE-Ireland: A User's Guide. Belfast: Cló Ollscoil na Banríona.

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

• • •

Sofia Rüdiger
University of Bayreuth
sofia.ruediger@uni-bayreuth.de

Preposition Overload? Examining the Use of Prepositions by Korean Speakers of English

Even though English is learned as a foreign language in South Korea it has a very profound status in the society: Not only is it recognized as the language of prestige and social advancement (see e.g. Park 2009) but it is also very visible within the country (cf. the influences of English on the Korean language as documented for example by Shim 1994). English in South Korea has so far mainly been investigated from a qualitative point of view and despite a plethora of studies on English language use on television, popular music, advertising and the linguistic landscape, next to a number of studies on language attitudes, large-scale investigations employing thorough corpus linguistic methods are still lacking (for a notable exception see Hadikin 2014). In this study, I employ the new Spoken Korean English Corpus to investigate the use of prepositions by Korean speakers of English. The corpus consists of 60 hours of spoken conversational material between 115 young educated Korean speakers of English and the researcher which as transcribed text amounts to ca. 300,000 words (excluding interviewer speech). The spoken nature of the corpus makes it ideal to observe trends and emerging patterns as "oral performance is less constrained and less conservative than written styles, so this is where innovations are most likely to surface" (Schneider 2004: 247).

In previous research, the overuse of prepositions has rather anecdotally been identified as a feature of a potential Korean English variety (Shim 1999). However, this study shows that variation in the use of prepositions by Korean English speakers is threefold, as 1) plus-prepositions, 2) minus-prepositions and 3) swap-prepositions (i.e. one preposition standing instead of another) can be found. Frequency-wise, minus-prepositions are the most commonly found variant in the corpus, followed by swap-prepositions. Plus-prepositions, the starting point of the investigation, are the least numerous in the data and, overall, seem to play only a minor role compared to the number of prepositions overall. Additionally, this study also identifies different contexts in which the previously mentioned variants of preposition use occur. For example, plus-prepositions occur as part of new multi-word verbs or in combination with a limited range of adverbs (e.g. *home*, *abroad*) whereas prepositions are most frequently omitted after verbs of movement such as *go* and *come* or when they are part of fixed multi-word verbs.

The results can be related to a number of insights from research in other contexts, as the innovative use of prepositions has been attested in the field of World Englishes (e.g. Sand 1999, Zipp 2014), English as a Lingua Franca (e.g. Cogo and Dewey 2012) and Second Language Acquisition (though there usually termed 'error' rather than 'innovation'; e.g. Alonso Alonso, Cadierno and Jarvis 2016).

References

Alonso Alonso, R., Cadierno, T. & Jarvis, S. (2016). Crosslinguistic influence in the acquisition of spatial prepositions in English as a foreign language. In: R. Alonso Alonso (Ed.), Crosslinguistic Influence in Second Language Acquisition. Bristol, Buffalo, Toronto: Multilingual Matters, 93–120.

Cogo, A. & Dewey, M. (2012). Analysing English as a Lingua Franca: A Corpus-driven Investigation. London: Continuum International Pub.

Hadikin, G. S. (2014). Korean English: A Corpus-driven Study of a New English. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Park, J. S.-Y. (2009). The Local Construction of a Global Language: Ideologies of English in South Korea. Mouton de Gruyter: Berlin/New York.

Sand, A. (1999). Linguistic Variation in Jamaica – A Corpus-Based Study of Radio and Newspaper Usage. Tübingen: Gunter Narr Verlag.

Schneider, E. (2004). How to trace structural nativization: particle verbs in world Englishes. World Englishes, 23 (2), 227-249.

Shim, R. J. (1994). Englishized Korean: Structure, status, and attitudes. World Englishes, 13 (2), 225-244.

Shim, R. J. (1999). Codified Korean English: Process, characteristics and consequence. World Englishes, 18 (2), 247–258.

Zipp, L. (2014). Educated Fiji English – Lexico-grammar and Variety Status. Amsterdam / Philadelphia: John Benjamins Publishing Company.

• • •

Karolina Rudnicka
University of Freiburg
karolina.rudnicka@frequenz.uni-freiburg.de

Sentence length, genres and grammatical obsolescence

The present paper uses a wide range of corpus data to test a widely held assumption that sentence length in written English has been steadily decreasing over the past few centuries. For more than a hundred years, empirical studies have shown that this claim is not entirely speculative, but at least partly based on linguistic fact. To the extent that sentence length is a correlate of syntactic complexity, it must be included in comprehensive studies of recent grammatical change.

More than one hundred years ago, Lewis (1894: 34) concluded that "the English sentence has decreased in average length at least one-half in three hundred years". In a much more recent study Schneider (2002) reports that the sentence length in newspaper news has decreased by an average of 15 words since 1700. Fries (2010) shows a decrease of approximately 10 words during the 18th century in the *London Gazette*,

while Westin (2002) provides statistically significant evidence for this effect in a wide range of English newspapers, such as *The Times*, *Guardian* and *Daily Telegraph*, in time period 1900-2000.

As can be seen, most research on sentence length has been based on newspaper data. The present paper, by contrast, offers a comprehensive analysis of the decrease in sentence length in the time period of 1800-2000, which is based on all the textual genres covered in the Corpus of Historical American English (COHA), i.e. *Fiction*, *Non-Fiction*, *Magazine* and *Newspaper* and one sub-genre – *Play & Script* (subpart of *Fiction*), which is used as proxy for spoken language.

The major linguistic aim of the paper is to relate changes in sentence length to changes in English syntax. The paper provides a proof of concept of the approach by using the decline in the frequency of the non-finite purpose subordinator *in order to* as a test case. Sentence length, sentence complexity and the likelihood of occurrence of *in order to* (rather than any of its less explicit variants) are shown to be interrelated in interesting ways. The immediate focus of the present paper is on a case of grammatical obsolescence. I will, however, argue that the systematic study of variation of sentence length across time and genre is also of interest to cognitive scientists, journalists, and language teachers.

References

Fries, U. (2010). Sentence Length, Sentence Complexity and the Noun Phrase in 18th-Century News Publications. In M. Kyto, J. Scahill, H. Tanabe (Eds.), Language Change and Variation from Old English to Late Modern English. Linguistic Insights 114. Bern: Peter Lang, 21-33.

Lewis, E. H. (1894). The History of the English Paragraph. The University of Chicago Press.

Schneider, K. (2002). The Development of Popular Journalism in England from 1700 to the Present, Corpus Compilation and Selective Stylistic Analysis. PhD Dissertation, Universitaet Rostock.

Westin, I. (2002). Language Change in English Newspaper Editorials. Language and Computers: Studies in Practical Linguistics, No. 44. Amsterdam: Rodopi

Davies, M. (2010-). The Corpus of Historical American English: 400 million words, 1810- 2009. Available online at http://corpus.byu.edu/coha/.

• • •

Chris Rühlemannn
Marburg University
chrisruehlemann@googlemail.com

Alexander Ptak
Paderborn University
aptak@mail.uni-paderborn.de

How long does it take to say *well*? Durations of words in context in CABNC

In this study we investigate the relationship between the durations, functions, and positions of words in conversational speaking turns. The study is based on CABNC (Albert et al. 2015), a collection of files from the British National Corpus (BNC) for which audio files are available (Coleman et al. 2012). What makes CABNC invaluable is the addition of measurements of the discourse durations of roughly 2 million words in the corpus. The measurements are recorded as attribute values in the XML structure of the corpus and can thus be extracted using XQuery, and analyzed statistically using R. Given the timings' sub-optimal accuracy rate (Renwick et al. 2013) we conducted not only large-scale quantitative but also in-depth qualitative analyses. The analyses were centred around the hypothesis that the duration of words increases over the speaking turn.

We examined more than 60,000 turns divided into subsets of turns ranging in length from three to ten words. We found two consistent patterns across all subsets. First, the duration of the turn-first item is greater than the duration of the turn-second item. Second, the durations of words see a consistent increase from the second position to the final position of the turn. To account for the two patterns we examined a random sample of nine-word turns and re-analyzed all the words' durations using Praat. The qualitative turn-by-turn analysis largely confirmed both the durational 'burst' in turn-first position and the consistent increase in durations over the turns. To account for the findings we discuss a number of explanatory hypotheses. They include an increase in phonemic size (cf. Zipf 1965), the end-focus maxim (Leech 1983), turn-final syllable lengthening (Levinson & Torreira 2015), and turn-end projection design. Finally, we also found significant differences in the duration of 'well' depending on its function and position in the turn, with speakers shortening 'well' used as a 'turn-preface' (Heritage 2015) while lengthening it when using it in a grammatical function as a an adverb.

The study may have important implications for the linguistic and conversation-analytic understanding of turn design in that timing of words seems to play a vital role in that design. Also, the study adds duration to the rich tableau of functional facets already discovered for the pragmatic marker 'well'.

References

Albert, S., de Ruiter, L. E. & de Ruiter, J. P. (2015). CABNC: the Jeffersonian transcription of the Spoken British National Corpus (available at: https://saulalbert.github.io/CABNC/)

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

Coleman, J., Baghai-Ravary, L., Pybus, J. & Grau, S. (2012). Audio BNC: the audio edition of the Spoken British National Corpus. Phonetics Laboratory, University of Oxford (available at: http://www.phon.ox.ac.uk/AudioBNC)

Heritage. J. (2013). Turn-initial position and some of its occupants. Journal of Pragmatics 57: 331-337.

Heritage, J. (2015). Well-prefaced turns in English conversation: A conversation analytic perspective. Journal of Pragmatics 88: 88–104.

Leech, G. (1983). Principles of pragmatics. London: Longman

Levinson, S. C. & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. Frontiers in Psychology, 6: 731. dos: 10.3389/fpsyg.2015.00731

Renwick, M. E. L., Baghai-Ravary, L., Temple, R. & Coleman, J. S. (2013). Assimilation of word-final nasals to following word-initial place of articulation in UK English, INTERSPEECH-2013, 3047-3051 (available at: http://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_3047.pdf)

Schegloff, E. A. (2007). Sequence organization in interaction: A primer in Conversation Analysis. Cambridge: Cambridge University Press.

Zipf, G. K. (1965). Human behavior and the principle of least effort. New York: Hafner

• • •

Alison Sealey
Lancaster University, UK
a.sealey@lancaster.ac.uk

Bob Carter
University of Leicester, UK
rc300@leicester.ac.uk

Corpora and assemblages – accommodating the non-human world

Corpus linguists are well aware that corpus **construction** entails making a series of decisions about categories – about what is to count as a word, a word class, a text-type, a genre, a speaker from a particular demographic group, or of a named variety, and so on.

Practices relating to classification are also evident in the **outputs** of corpus analysis. Corpus-assisted discourse analysis highlights, for example, patterns in how people are represented as belonging to particular categories (e.g. 'refugees'), and differentially depicted (e.g. as 'floods') (Gabrielatos & Baker 2008). Indeed, it is often claimed that the discourse 'constructs' the objects it names.

This presentation explores issues of categories and classification with reference to discourse, not about human groups but rather about the non-human, using evidence from a specialised corpus comprising texts that are about non-human animals[1]*. A sub-set of this corpus, of approximately 6 million words, comprises articles featuring a wide range of kinds of animal, published in journals concerned with biology, veterinary and environmental science etc. We used a bespoke script in R that enabled us to tag every instance of animal naming terms with a specific symbol so as to retrieve them for analysis. This facilitated identification of the most frequently named animals, while further analysis, using *AntConc*, *TagAnt* and *Sketch Engine* (Anthony 2015), was of the linguistic means by which these animals are depicted variously as integral beings, as components of larger collectivities, and as anatomised and segmented. Syntactic and phraseological patterns identified through corpus analysis also reveal how these animals – and parts of their bodies – are represented as featuring in human enterprises.

This corpus assisted discourse analysis draws on a theory in which there is a growing interest within the social sciences. 'Assemblage theory', which derives primarily from the work of Deleuze and Guattari, has recently been developed and modified, principally by Manuel DeLanda. Whereas the theories on which the social sciences conventionally draw tend to have an implicit anthropocentric bias, this is less the case with the ontology of assemblage theory, which assumes the heterogeneity of all assemblages and the importance of relations between them. In the case of living organ-

1 This research project *'People', 'products', 'pests' and 'pets': the discursive representation of animals* is funded by the Leverhulme Trust (RPG 2013-063)

isms, these relations include the material resources and processes that enable them to grow and flourish (air to breathe, food to eat etc.), even while they move towards death and decomposition. In the case of language, DeLanda (2016) proposes that 'linguistic entities operat[e] as variables of a collective assemblage'. Thus the concept of assemblages, at different scales, is applicable across the range of complex and ever-changing phenomena that exist in the physical, natural, social - and semiotic - world.

In this paper, we aim to demonstrate how viewing our corpus findings from an assemblage perspective can lead to new insights into the linguistic construal of the natural world.

References

Anthony, L. (2014). AntConc (Version 3.4.3). Tokyo, Japan. Retrieved from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2015). TagAnt (Version 1.2.0). Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/

DeLanda, M. (2016). Assemblage Theory. Edinburgh: Edinburgh University Press. p.64

Deleuze, G., & Guattari, F. (1983). Anti-Oedipus. Minneapolis: University of Minnesota Press.

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. Journal of English Linguistics 36 (5), 5 - 38.

• • •

Ole Schützler

University of Bamberg

ole.schuetzler@uni-bamberg.de

Grammaticalization and subjectification as correlated processes? Diachronic developments of the concessive prepositions *in spite of* and *despite* in British and American English

Building on work by Rissanen (2002), this paper traces developments of the concessive prepositions *in spite of* and *despite* in British and American English from the early seventeenth century to the present day. The leading questions are: (i) When did *despite* become more frequent than *in spite of*, (ii) did the changes begin earlier or proceed more rapidly in BrE or AmE, and (iii) were changes in frequency accompanied by changes in semantic or pragmatic functions? Analyses are based on corpus data from

ARCHER (cf. Yáñez-Bouza 2011), COHA (Davies 2010–), the extended Brown family of corpora (c.f. Baker 2009), and the Hansard Corpus (Alexander & Davies 2015–).

There is a steady increase in the frequency of both prepositions up until the middle of the twentieth century, which then continues and speeds up in *despite*, while *in spite of* begins to decline again. This shift in preference is argued to result from a process of specialization in ongoing grammaticalization within the domain of concessive prepositions (e.g. Hopper 1991). The change happens earlier and is more dynamic in American English. Over the last 150 years, both prepositions were very predominantly used in so-called 'content concessives' (Sweetser 1990), in which there is some presupposed real-world negative conditional or causal relationship between propositions – a so-called topos (cf. Azar 1997) – that remains unrealized. In (1), the topos is based on the expectation that an agile person will normally (but not in this case) be able to escape from a dangerous situation and save himself.

(1) Despite the remarkable agility of the Sioux, he could not save himself […]. (COHA, 1889, fiction)

However, the data reflect a slow-moving but clearly directional semantic change, whereby *in spite of* and *despite* become more likely to be used in so-called 'speech-act concessives' (again cf. Sweetser 1990), as illustrated in (2): One of the two propositions contrasts with, corrects or qualifies the other, not based on a clearly identifiable topos but by taking an alternative pragmatic stance. Such constructions are often argued to be more subjective, since the proposition carrying the concessive marker is not an 'objective', generally understood real-world obstacle to the main-clause proposition but modifies the speech-act presented there (cf. Crevels 2000). In this case, simultaneously reporting a smiling expression and a serious tone of voice suggests two conflicting interpretations of a single situation.

(2) In spite of Jim's smiling way, there was something serious-sounding in his voice […]. (ARCHER, AmE, 1948, fiction)

Concessives of another type identified by Sweetser, so-called 'epistemic concessives', are rather rare in the data and seem to play a marginal role at best.

The evidence suggests that there is indeed what Traugott (1995) calls subjectification in grammaticalization, i.e. subjectification concomitant of grammaticalization. However, the semantic change progresses much more slowly than frequency changes, which were interpreted as symptoms of grammaticalization. The two processes may be related, but they do not seem to be tidily aligned; rather, subjectification follows grammaticalization.

References

Alexander, M. & Davies, M. (2015–). *Hansard Corpus 1803–2005*. Available online at http://www.hansard-corpus.org.

Azar, M. (1997). Concessive Relations as Argumentations. *Text* 17(3), 301–316.

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics* 14(3), 312–337.
Crevels, M. (2000). Concessives on different semantic levels: A typological perspective. In: E. Couper-Kuhlen & B. Kortmann (eds.), *Cause – Condition – Condition – Contrast. Cognitive and Discourse Perspectives*. Berlin: Mouton de Gruyter. 313–339.
Davies, M. (2010–). *The Corpus of Historical American English: 400 million words, 1810–2009*. Available online at http://corpus.byu.edu/coha/.
Hopper, P. J. (1991). On some principles of grammaticization. In Elisabeth Closs Traugott & Bernd Heine (eds.), *Approaches to Grammaticalization*. Vol. I: *Focus on Theoretical and Methodological Issues*. Amsterdam: John Benjamins. 17–35.
Rissanen, M. (2002). Despite or notwithstanding? On the development of concessive prepositions in English. In A. Fischer, G. Tottie & H. M. Lehmann (eds.), *Text Types and Corpora; Studies in Honour of Udo Fries*. Tübingen: Gunter Narr. 191–203.
Sweetser, E. E. (1990). *From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.
Traugott, E. Closs. (1995). Subjectification in grammaticalisation. In D. Stein & S. Wright (eds.), *Subjectivity and Subjectivisation. Linguistic Perspectives*. Cambridge: Cambridge University Press. 31–54.
Yáñez-Bouza, N. 2011. ARCHER past and present (1990–2010). *ICAME Journal* 35, 205–236.

• • •

Martin Schweinberger
University of Kassel
martin.schweinberger.hh@gmail.com

Using intensifier-adjective collocations to determine mechanisms of change

This study takes a corpus-based approach to examining co-occurrence patterns of amplifying intensifiers and adjectives (cf. 1) based on two components of the International Corpus of English (ICE). From a language variation and change perspective intensifiers are particularly interesting as they play a crucial part in the "social and emotional expression of speakers" (Ito & Tagliamonte 2003: 258) and because intensifier systems are prone to change (Ito & Tagiamonte 2003:257; Quirk et al. 1985:590).

The study of bigram collocations, i.e. co-occurrence patterns, of intensifiers and adjectives utilizes data from two distinct regional varieties of English, New Zealand English (NZE) and Irish English (IrE), which differ markedly with respect to ongoing changes in their intensifier systems: while VERY is receding in both varieties,

REALLY is increasing in frequency in New Zealand English in apparent-time leading D'Arcy (2015) to hypothesize that REALLY is replacing VERY , while REALLY is not increasing in apparent-time in Irish English thus showing that REALLY is not replacing VERY.

(1) a. It's a very elegant technique (ICE NZ: S2A-038)
b. oh wow that's really cool (ICE NZ: S1A-096)
c. she looks bloody old in that picture any rate (ICE NZ: S1A-096)

The assumption underlying the present study is that lexical replacement requires two variants to exhibit similar collocational profiles while this should not be the case if the processes of increase and decline in apparent-time are independent.

Given the two distinct trajectories of REALLY in NZE and IrE, this would predict that REALLY and VERY show very similar collocational profiles in NZE while their profiles in IrE should be distinct.

To extract all adjectives, the corpus data was POS-tagged by implementing a maximum entropy part-of-speech tagger. For each adjective, it was determined whether or not is was intensified and which type of intensifier occurred. The statistical analysis applied the principle of accountability and used correspondence and cluster analyses to investigate the above stated hypotheses by determining similarities between collocational profiles of intensifiers and adjectives. Additional configural frequency analyses (CFAs) were used to determine whether co-occurrence patterns were associated with specific age groups as the co-occurrence patterns among younger speakers should differ from the patterns observed among older speakers.

The results of the statistical analyses show that the co-occurrence patterns of REALLY and VERY in NZE are indeed similar thus indicating lexical replacement while the co-occurrence patterns of REALLY and VERY in IrE are distinct which would explain why REALLY is not successfully replacing VERY in this regional variety.

The similarity of the co-occurrence pattern of VERY and REALLY in NZE pose the question of whether there is a more general, cross-varietal trend at work in lexical replacement which underlies the restructuring of the intensification system across varieties of English.

References

Ito, R. & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. Language in Society 32: 257—279.

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

D'Arcy, A. F. (2015). Stability, stasis and change – the longue durée of intensification. Diachronica 32(04), 449–493.

• • •

Gries Stefan Th.
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

Wahl Alexander
Radboud University
awahl1@gmail.com

MERGE: A new recursive approach towards multiword expression extraction and four small validation case studies

In this paper, we introduce a new bottom-up approach to the identification/extraction of multi-word expressions (MWEs) in corpora. This approach, Multi-word Expressions from the Recursive Grouping of Elements (MERGE), is based on the successive merging of bigrams to form word sequences of various lengths; the selection of bigrams to be ‚merged' is at present based on the use of the log likelihood measure but other association measures can also be used. We discuss how the method is applied to corpus data and, to test/validate its performance, we discuss four different validation studies.

First, we applied the algorithm to the Santa Barbara Corpus of Spoken American English and ICE-Canada to identify MWEs that, according to MERGE, are ‚good' and ‚worse' MWEs and then had native speakers of American English rate them on a Likert scale for how much the MWE stimuli constituted „a complete unit of vocabulary" to see whether the native speakers reproduced the results of our algorithm. We analyzed the ratings with a linear mixed-effects model that also accounted for the MWEs' lengths and found that MERGE's output indeed distinguishes significantly between ‚good' and ‚worse' MWEs.

Second, we compared the output of MERGE when applied to the above corpora to the output of Brook O'Donnell's Adjusted Frequency List by having (different) native speakers rate them on a Likert scale again; a linear mixed-effects model found that the MWEs returned by MERGE score significantly higher than those returned by the AFL.

Third, we applied both MERGE and AFL to the complete spoken component of the BNC to determine how well both methods can identify 388 expressions that the compilers of the BNC decided to tag as multi-word units (using the <mw></mw> tag). We took the top 10,000 items from either approach and used one-tailed binomial tests to see whether MERGE would perform better than AFL on this task (given the previous case study); testing in both directions (against both baselines), we found that MERGE found significantly more BNC multi-word units than the AFL, and the AFL found significantly fewer BNC multi-word units than MERGE.

Finally, we explored MERGE's performance using L1-acquisition corpus data. We ran MERGE on both the adult and the child utterances of the Lara and Thomas corpora from the CHILDES collection. Specifically, we split adult and child data into an early (the first $^{2}/_{3}$) and a late part (the last $^{1}/_{3}$) and compared the MERGE scores of the adult MWEs that the children used in the late partition to the MERGE scores of the adult MWEs that the children did not use later. A linear model again controlling for

both the MWE lengths and which child's data we were exploring found a significant interaction of MWE length, MWE strength of association, and child, but showed that MWEs with higher MERGE scores are indeed those that children are more likely to learn even when length is controlled (with slight differences across children).

We conclude by integrating all results, discussing their implications, and suggesting future analyses.

• • •

Hang Su
Beihang University
suhangunique@hotmail.com

Developing local grammars: A pattern-based approach

Local grammar research has been gaining increasing popularity in the linguistic community (e.g. Hunston & Sinclair 2000; Barnbrook 2002; Bednarek 2008; Su 2015, 2017; Cheng & Ching 2016; Warren & Leung 2016). This tradition of research has both theoretical and practical significance. Theoretically, the exercise of local grammar analysis itself can offer insights into the relationship between lexis, meaning, and grammar; and practically, local grammars developed for each meaning area can be used to assist automatic extraction of units that are associated with that meaning (Barnbrook & Sinclair 2001; Bloom 2011).

Currently, research on local grammars has mainly focused on one particular meaning or discourse function; for example, efforts have been made to develop local grammars of evaluation (Hunston & Sinclair 2000; Hunston & Su forthcoming), definition (Barnbrook 1995, 2002), disclaiming (Cheng & Ching 2016), and so on. These studies are certainly valuable because they contribute significantly to a comprehensive description of the chosen meaning or function. The problem, however, is that such research is restricted in the sense that it only contributes to local grammars of this and/or that. In addition, since the local grammars were developed by different researchers, the analytic methods and coding systems used in these studies differ sharply from one another. This then raises the question as to how can local grammars be developed more systematically and consistently?

This paper addresses this question and proposes an approach that takes grammar patterns as its starting point to develop local grammars. The rationale behind this approach is the observation that patterns and meanings are associated (Hunston & Francis 1999). Specifically, since words occurring in a pattern can be classified into a limited set of meaning groups, each pattern can be analysed using a limited set of local grammar patterns, with each set of local grammar patterns accounting for

one meaning group. This paper, then, reports a preliminary study of 5 verb patterns selected from Francis, Hunston and Manning (1996), aiming to show the feasibility of using a pattern-based approach to develop local grammars. These patterns are **V n to-inf.**, **V n** *that*, **V n** *about* **n**, **V n** *in* **n**, and **V n** *of* **n.**

Using examples from Francis et al (1996) and the Bank of English, the analyses show that instances of each pattern can be analysed using a few functional elements associated respectively with each meaning. Provided that all patterns have been analysed, local grammar patterns of each meaning can then be generalised and grouped together. In doing so, cumulative coverage of the description of different meanings, and ultimately a generalisation of semantic descriptions, can be achieved. In this sense, it can be argued that a pattern-based approach is useful for developing systematically and consistently a set of local grammars that can be used to account adequately for language in use.

References

Barnbrook, G. (2002). Defining Language: A Local Grammar of Definition Sentences. Amsterdam: John Benjamins.

Barnbrook, G. & Sinclair, J. (1995). Parsing Cobuild entries. In John Sinclair, Martin Hoelter & Carol Peters (eds.), The Language of Definition: The Formalization of Dictionary Definitions for Natural Language Processing, 13–58. Luxembourg: European Commission.

Barnbrook, G. & Sinclair, J. (2001). Specialised corpus, local and functional grammars. In Mohsen Ghadessy, Alex Henry & Robert Roseberry (eds.), Small Corpus Studies and ELT: Theory and Practice, 237–276. Amsterdam: John Benjamins.

Bednarek, M. (2008). Emotion Talk across Corpora. New York: Palgrave Macmillan.

Bloom, K. (2011). Sentiment Analysis Based on Appraisal Theory and Functional Local Grammar. PhD thesis. Illinois Institute of Technology, US.

Cheng, W & Ching, T. (2016). 'Not a guarantee of future performance': The local grammar of disclaimers. Applied Linguistics. DOI:10.1093/applin/amw006.

Francis, G., Hunston, S. & Manning, E. (1996). Collins Cobuild Grammar Patterns 1: Verbs. London: Harper Collins.

Hunston, S. & Francis, G. (1999). Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English. Amsterdam: John Benjamins.

Hunston, S. & Sinclair, J. (2000). A local grammar of evaluation. In S. Hunston & G. Thompson (eds.), Evaluation in text, 74–101. Oxford: Oxford University Press.

Su, H. (2015). Judgement and Adjective Complementation Patterns in Biographical Discourse: A Corpus Study. PhD thesis. University of Birmingham, UK.

Su, H. (2017). Local grammars of speech acts: An exploratory study. Journal of Pragmatics. 111: 72–83.

Warren, M. & Leung, M. (2016). Do collocational frameworks have local grammars? International Journal of Corpus Linguistics 21(1), 1–27.

• • •

Sali A. Tagliamonte
University of Toronto
sali.tagliamonte@utoronto.ca

Bridget Jankowski
University of Toronto
bljankowski@gmail.com

Golly, Gosh and *Oh my God!* What dialect corpora can tell us about swearwords

According to 18th century doctrine, the name of God should be used sparingly and reverently, without familiarity, and not used in daily life (Gibson 1760:7). In subsequent centuries the same words came to be used to intensify utterances (Joseph 2006:134), and in the 20th century the use of swear words generally has become championed as features of self-expression and human bonding (e.g. Adams 2016). Among innumerable swear words, the use of expressions and euphemisms of "God", henceforth "G-words," are particularly colourful.

In this study, we conduct a comparative cross-variety investigation of G-words, as in (1-5), in a large compendium of dialect corpora from Ontario, Canada.

1. *By gee,* we thought it was crazy in a way you-know but *by golly* it paid off. (J. Mcleech, b. 1899)
2. *Geez* like with that Canadian accent. (J. Ponce, b. 1984)
3. Well, *Christ*, we'll freeze to death in this damn house. (B. Leblanc, b. 1937).
4. *Oh God* it was cold up there. (C. Winter, b. 1924)
5. *Oh my God*, it was just too expensive. (B. Mastersen, b. 1961)

The dialects contrast by urban vs. rural, population size, and distance from a large urban centre, permitting insight into this variation across geographic space. Because the corpora are generationally (individuals born between 1879-1999) and sociolinguistically stratified, we can examine change in apparent time and according to social factors. Moreover, with N= 2286 tokens of innumerable G-words, there is sufficient data for comparative sociolinguistic techniques and statistical modelling (e.g. mixed effects regression).

The results expose a number of striking findings. Using apparent time as a proxy for historical change, we discover that G-words have undergone remarkable shift across the 20th century. Rural communities show retention of euphemisms such as *golly, gee(z),* and *gosh,* and there are notable social contrasts to their use: Females favour *gosh* while males favour *gee(z),* and euphemisms are used more by less-educated speakers. Variants with *God* are not only predominant in the urban centre, they seem to encompass societal change across the whole territory. Younger speakers in every community shift towards non-euphemistic practise beginning in the 1930's and increasing after WWII. However, this shift is not entirely lexical replacement. Where

once individuals used *God* in collocations such as *Praise God* or *Thank God,* people born in the early 1960's onwards are using *God* in one collocation in particular: *Oh my God* (N=611). A fascinating correlate is that, as with many changes, this is being lead by higher-educated women who have white-collar jobs. This suggests that there has been a change towards a greater acceptance of the actual 'G-word' itself in contemporary society.

Taken together, these results highlight how synchronic dialect data reveal linguistic adjustments as well as cultural adaptations. Once "enlisting God in our verbal behavior" (Adams 2016:23) was only for oath-taking with euphemisms gradually evolving for expressive purposes. However, in the late the 20th century, euphemisms are no longer preferred. Instead, *God* is used indiscriminately as a secularized expression of emotional intensity.

References

Adams, M. (2016). In praise of profanity. Oxford: Oxford University Press.

Joseph, J. (2006). Language and Politics. Edinburgh: Edinburgh University Press.

Gibson, E., The Right Rev., Lord Bishop of London. (1760). Admonition against Profane and Common Swearing: in a Letter from a Minister to his Parishoners; to be put privately into the Hands of Persons who are addicted to Swearing, 20th edn, London: printed by E. Owen in Warwick-Lane, and sold by W. Johnston in Ludgate-Street.

• • •

Laura Terassa
University of Freiburg, Germany
laura.terassa@frequenz.uni-freiburg.de

Past tense and plural omission as potential means to avoid redundancy in Singapore English: Insights from corpus data and a perception experiment

According to Trudgill (2001: 372-373), loss of redundancy reduces complexity in language contact settings. For the contact variety Singapore English (SgE), it has been stated that the past time reference is often indicated by means of time adverbials and not through verbal inflection in colloquial speech (e.g. Bao 1998: 163). Inflectionally unmarked nouns with plural reference have been shown to occur both in the presence (e.g. Wee & Ansaldo 2004: 64) and in the absence (e.g. Ho & Platt 1993: 22) of quantifiers (compare also Ziegeler 2015: 182-183).

This paper investigates the impact of preceding time adverbials with past time reference on lack of inflectional past tense marking in verbs and the impact of preceding quantifiers with plural reference on lack of inflectional plural marking in nouns for SgE. It is hypothesized that preceding time adverbials and quantifiers trigger the production and facilitate the perception of inflectionally unmarked verbs and nouns among SgE speakers because they make the inflectional affixes redundant.

As production data, the spoken parts of ICE Singapore were used. For random samples of 20 regular verbs ending in a vowel and 20 nouns, the degrees to which the target forms lack their inflectional affixes were identified in general and in the presence of a preceding time adverbial and quantifier in particular. Only verbs ending in a vowel were chosen to exclude final consonant cluster reduction as an explanation for the missing *-ed* suffix. The perception of the same sample of verbs and nouns in the presence and absence of a time adverbial or quantifier was tested by means of a web-based perception experiment that comprised a self-paced reading task and an acceptability judgment task. Speakers of American English (AmE) and British English (BrE) served as the control group.

The corpus analyses reveal low rates of lack of past tense marking (compare Gut 2009: 266) and plural marking (compare Deterding 2007: 42-44) and show that the majority of the unmarked verbs and nouns is not preceded by a time adverbial or quantifier, respectively. For the experimental data, linear mixed effects models indicate that speakers of SgE read verbs that lack inflectional past tense marking (and the words directly following them) faster than the control group, irrespective of whether a time adverbial precedes the unmarked verb or not. They also judge the unmarked verbs better (particularly when a time adverbial precedes them). Nouns that lack inflectional plural marking are read faster and judged better by the SgE speakers than by the control group, when they are not preceded by a quantifier with plural reference.

The findings indicate that preceding time adverbials and quantifiers do not trigger the production or facilitate the perception of the inflectionally unmarked verbs and nouns in the sample to the expected degree. Irrespective of their presence, past tense and plural omission is relatively rare but salient enough to be familiar to speakers of SgE.

References

Alsagoff, L., & Ho, C. L. (1998). The grammar of Singapore English. In J. A. Foley et al. (Eds.), English in new cultural contexts: Reflections from Singapore. Singapore: Singapore Institute of Management, 127-151.

Bao, Z. (1998). The sounds of Singapore English. In Joseph A. Foley et al. (Eds.), English in new cultural contexts: Reflections from Singapore. Singapore: Singapore Institute of Management, 152-174.

Deterding, D. (2007). Singapore English. Edinburgh: Edinburgh University Press.

Gut, U. (2009). Past tense marking in Singapore English verbs. English World-Wide, 30 (3), 262-277.

Trudgill, P. (2001). Contact and simplification: Historical baggage and directionality in linguistic change. Linguistic Typology, 5 (2), 371-374.

Wee, L., & Ansaldo, U. (2004). Nouns and noun phrases. In L. Lim (Ed.), Singapore English: A grammatical description. Amsterdam: John Benjamins, 57-74.
Ziegeler, D. (2015). Converging grammars: Constructions in Singapore English. Berlin: De Gruyter Mouton.

• • •

Faye Troughton
Université de Mons
faye.troughton@umons.ac.be

Lobke Ghesquière
Université de Mons
lobke.ghesquiere@umons.ac.be

What a transformation! On *what*'s development from identity to degree

This paper reports on a diachronic data study of *what*, charting both its identifying (1) and degree readings (2). In doing so, this paper aims to verify two hypotheses available in the literature, i.e. Bolinger's (1972) classic shift from identification to intensification and Siemund's (2015) recently posited gradient of clause types.

(1) *What tools would be required?* (CLMET3_0_3_312.)
(2) *What a delightful surprise*! (WB brbooks)

Bolinger (1972: 61) hypothesized on the basis of synchronic observation a productive pathway of change for elements of the English noun phrase leading from identification to intensification, i.e. from determiner to degree modifying functions. This shift has already been examined in detail and confirmed for *such*, both synchronically (e.g. Mackenzie 1997) and diachronically (e.g. Ghesquière & Van de Velde 2011). Despite Bolinger's (1972: 91) assertion that the interrogative determiner *what* has "undergone a similar shift in an extensible sense", this has yet to be verified. So far, focus has generally been on the determiner function of *what*, with its intensifying uses having only been studied synchronically and their development speculated upon. A potential reason for this gap in the literature becomes clear upon looking into any comprehensive grammar, as *what's* intensifying function is rarely recognised outside of its role in the exclamative clause, the prime syntactic expression of extreme degree, as in "*what a wimp!*". Degree readings of other clause types involving *what*, as demonstrated in (3) and (4), are rarely included in the discussion.

(3) *Grace, do you hear* **what heresy** *Fanny has been learning***?** *Why, the proportion of ozone in the air here has been calculated to be five times that of even Aveton!* (CLMET3_0_3_197)

(4) *…houses much like his own, except that…they were terraced and* **what gardens** *they had were hidden by high walls* (WB brbooks)

The principal aim of this study is to verify if *what* follows the proposed pathway of development, from identifier to intensifier. This will be achieved through a diachronic data study of written British English, from 1500 to the present day, using the Penn-Helsinki Parsed Corpora of Middle English and Early Modern English, the Corpus of Late Modern English Texts 3.0, and the British books section of the Wordbanks Online Corpus. For each 70-year period 300 random examples were analysed. Parameters taken into account include collocational behaviour, clause type and direct vs. indirect speech.

Through an analysis of the clause types in which *what* expresses degree, this study also sets out to verify Siemund's (2015: 723) cline of clause types, a gradient of major to minor clause types, rather than the traditional "positions of binary opposition". Major clause types, such as the declarative, can convey almost any illocutionary force, whereas those at the other end of the cline, such as the exclamative, are said to be "considerably more restricted functionally" (Siemund 2015: 723). This gradient concept supports currently accepted restrictions placed on exclamative clauses, namely that they can only receive a degree reading (Rett 2011: 417) and are severely limited in their potential syntactic structure. The gradient also however allows for more flexibility and pragmatic use of language. As well as testing Siemund's cline in terms of the availability of degree readings across clauses, this paper tests the validity of the current focus on intensifying *what* as a part of the inherently evaluative exclamative clause, allowing us to evaluate *what*'s role as a degree modifier outside of this context.

References

Bolinger, D. (1972). Degree Words. The Hague: Mouton.

Ghesquière, L. & Van de Velde, F. (2011). A corpus-based account of the development of English ‚such' and Dutch ‚zulk'. Cognitive Linguistics, 22(4), 765–797.

Kennedy, C. & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. Language, 81(2), 345–381.

Mackenzie, J.L.M. (1997). Grammar, discourse and knowledge: The use of ‚such'. In J. Aarts, I. De Mönnink & H. Wekker (eds.), Studies in English language and teaching: In honour of Flor Aarts. Amsterdam: Rodopi, 85–105.

Rett, J. (2011). Exclamatives, degrees and speech acts. Linguist and Philos, 34, 411–442.

Siemund, P. (2015). Exclamative clauses in English and their relevance for theories of clause types. Studies in Language, 39(3), 697–727.

Trotta, J. (2000). Wh- Clauses in English: Aspects of theory and description. Amsterdam: Rodopi.

• • •

Susanne Wagner
JGU Mainz
susanne.wagner@uni-mainz.de

Amplifier-adjective 2-grams world-wide: focus on *pretty*

Among the most prominent means of intensifying in English are amplifiers modifying adjectives. For the past few decades, *very* (1), *really* (2) and *so* (3) seem to have shared the top slots in terms of relative frequency. *Pretty* (4) is a relative newcomer, mostly associated with American English (Biber et al. 1999).

(1) when they're going off to relax in their *very* **nice** cottages in the countryside (GH G)
(2) It makes you feel *really* **cool** and powerful! (AU G)
(3) *So* **glad** this collection has been preserved . (US B)
(4) and im *pretty* **sure** you hav no proof at all that the earth is 4.3b years old? (*sic*; HK G)

Previous studies have analysed the historical development of amplifiers (Gonzalez-Diaz 2008, Mendez-Naya 2008), preferences according to genres and (major) varieties (Biber et al. 1999). Most attention has been paid to variation and change in the system and the role of social factors in it, where notions such as "recycling" (Ito & Tagliamonte 2003, Tagliamonte 2008) and rapid change (Barnfield & Buchstaller 2010, Macauley 2006, Tagliamonte & Roberts 2006) are emphasised.

This paper uses data cumulated from GloWbE (Davies 2013; cf. Davies & Fuchs 2015) to contrast and compare six major English-speaking regions (US, GB, Australia/New Zealand, Indian subcontinent, South East Asia, Africa) with regard to

(1) their distributions and preferences concerning amplifier use, including statistical analysis for all top 10 amplifier-adjective pairs
(2) their preferred amplifier-adjective pairs
(3) the collostructional/collexeme status (Gries & Stefanowitch 2004) of these pairs vis-à-vis each other and for same pairs across varieties
(4) the impact of adjective frequency on amplifier-adjective collexeme status.

Overall distributional preferences show some regional patterning (e.g. the same four amplifiers form the top 4 in the same order in all regions), but do not confirm earlier assumption regarding e.g. the spread of *so* (Tagliamonte & Roberts 2006). Rather, they indicate that results should not be generalised across amplifiers or adjectives, but maximally across amplifier-adjective pairs (examples (1) to (4) represent some typical 2-grams). A comparison of low-, mid- and high-frequency adjectives in amplified contexts shows that increased adjective frequency leads to increased cueness/collexeme status. Collexeme analysis of 150 amplifier-adjective pairs also discloses broad international similarities but also regionally distinctive sub-patterns.

The status of *pretty* as an amplifier differs considerably around the world: for one, *pretty* has clearly left behind its US origins and has spread both regionally and semantically, combining with an increasing number of absolute adjectives (such as *perfect, unique*). Interestingly, however, although certain regions are clearly not *pretty* territory, certain 2-grams nevertheless share outlier status world-wide (primarily *pretty good*). Results presented in this paper not only add to our knowledge of amplifier use in (varieties of) English, but also indicate that it is time to look beyond the individual word and into larger units such as 2-grams.

• • •

Sebastian Wagner
Augsburg University
sebastian.wagner@philhist.uni-augsburg.de

Cromwell's power had never been so immensely great as at this very moment - The discoursal function of graduation in LModE historiography

Historical events, personae and their actions are the targets of evaluative assessment that are selected on the basis of what historiographers consider significant with regard to conveying their respective conception of historical knowledge. This prioritising and interpretative facet of history writing finds expression in the works composed in the LModE period and can be rendered visible on the interpersonal level: Initially, it was deemed fundamental by many 18th century historians to teach readers moral lessons by overtly presenting their judgments. Scholars of the emerging 'scientific' historiography considered it more important to convince readers of the plausibility of their interpretations (cf. Hesketh 2011, Martin 2002). In each case the underlying linguistic realisations of evaluative meaning reflect the historiographers' claim for interpretational sovereignty.

In contrast to the extensive corpus-based research on explicit manifestations of stance in academic discourse (e.g. Biber 2006, Hyland 2005), the genre of history writing still lacks quantitative studies concentrating on the (strategic) use of evaluative language. The purpose of this paper is to explore the discoursal functions of lexicogrammatical resources that can be categorised with reference to GRADUATION (Martin & White 2005). GRADUATION is a sub-system of the SFL-based Appraisal framework, which Goźdź-Roszkowski and Hunston (2016:134) consider "probably the most fully theorised view on evaluative language". The process of grading evaluative items can be classified by reference to "two axes of scalability", FORCE ('intensity') and FOCUS ('prototypicality') (Martin & White 2005:137). This categorisation is ex-

pected to not only permit quantification of scalable evaluative meaning across the different historiographical documents but also to allow for an observation of recurring (genre specific) patterns.

Preliminary findings taken from the 'Corpus of Late Modern British Historical Writing' (ca. 1,5m words), which comprises the excerpts of the works of 50 historians, published between 1700 and 1914, suggest a diversified utilisation of grading devices: In (1) the author marks the significance of the proposition by modifying the adjective with a maximizer (Quirk et al 1985:589). Here, the up-scaling to the highest possible intensity might reveal the attempt to align a putative reader with the historiographer's self-proclaimed interpretative authority. In example (2), almost functions so as to weaken the strength of the negatively charged evaluative phrase. By doing this, the historian confronts the reader with both the recounted and the hypothetical alternative outcome. The retrospective evaluation of historical entities made at the time of the discourse production is characteristic for historiography (cf. Bondi & Mazzi 2009): The up-scaling through highly in (3) illustrates this practice of highlighting new and particularly relevant insights for the reader.

(1) It is however *extremely* remarkable, that… (Belsham 1805)
(2) in the attack of fort Lazaro the English were repulsed with an *almost* incredible loss (Bigland 1813)
(3) A knowledge of that tongue must therefore have been *highly* useful… (Freeman 1872)

References

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

Bondi, M. & Mazzi, D. (2009). Writing history: argument, narrative and point of view. In M. Shiro; P. Bentivoglio & F. Erlich (Eds.), Haciendo discurso. Homenaje a Adriana Bolívar. Caracas: Universidad Central de Venezuela, 611-626.

Dew, B. & Price, F. (2014). Introduction: Visions of History. In: B. Dew & F. Price (Eds.), Historical writing in Britain, 1688-1830. Springer, 1-17.

Feldner, H. (2010). The new scientificity in historical writing around 1800. In: S. Berger, H. Feldner, K. Passmore (Eds.), Writing History: Theory and Practice, London: Hodder Arnold H&S, 3-21.

Goźdź-Roszkowski, S. & Hunston, S. (2016). Corpora and beyond–investigating evaluation in discourse: Introduction to the special issue on corpus approaches to evaluation. Corpora, 11(2), 131-141.

Hesketh, I. (2011). The Science of History in Victorian Britain: Making the Past Speak (Science & Culture in the Nineteenth Century). New York/London: Routledge.

Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. Discourse Studies, 7 (2), 173-191.

Martin, J. R. (2002). Writing History: Construing Time and Value in Discourses of the Past. In: M. J. Schleppegrell & C. M. Colombi (Eds.), Developing Advanced

Literacy in First and Second Languages Meaning with Power, New Jersey: Lawrence Erlbaum Associates, 87-118.
Martin, J. R. & White, P. R. R. (2005). The Language of Evaluation: Appraisal in English. Basingstoke: Palgrave.
Quirk, R. et al. (1985). A comprehensive grammar of the English language. London: Longman.

• • •

Viola Wiegand
University of Birmingham
wiegandv@bham.ac.uk

Anthony Hennessey
University of Nottingham
anthony.hennessey@nottingham.ac.uk

Christopher R. Tench
University of Nottingham
christopher.tench@nottingham.ac.uk

Michaela Mahlberg
University of Birmingham
m.a.mahlberg@bham.ac.uk

Comparing co-occurrences between corpora

The concept of 'collocation' is one of the most fundamental ones in corpus linguistics. Although corpus linguistic approaches are inherently comparative, only a few studies have taken the analysis and interpretation of co-occurrence patterns beyond a single corpus. For example, Gabrielatos and Baker (2008) identify 'consistent collocates' across annual newspaper subcorpora describing refugees and asylum seekers. Baroni and Bernardini (2003) focus on collocational difference to compare the extent of collocation between Italian texts and texts translated into Italian. The Sketch Engine Word Sketch function (Kilgarriff et al. 2014) offers a comparison of the collocates of two words between subcorpora (available from www.sketchengine.co.uk). Examples like these highlight the need for formal statistical methods for comparing co-occurrence patterns between corpora.

In this paper, we propose a systematic method to support the analysis of differences in the collocational behaviour of words across corpora. The method identifies statistically significant co-occurrence count differences between two corpora and reports an effect size and confidence interval for each of the identified differences. We

explain the overall methodological approach and give additional detail of important statistical concepts; for example, how a rigorous comparison of co-occurrences requires the consideration of the false discovery rate when carrying out multiple statistical tests. We discuss the challenge of presenting and interpreting the significant results and suggest suitable visualisations and possible pitfalls. Both the method and visualisation tools are made available in an R package and we briefly discuss the implementation using R.

The method we propose will be illustrated using a case study comparing novels by Charles Dickens to other 19th century novels. The focus is on collocations that have been identified as literary relevant collocations for the description of fictional people. This case study is situated within the theoretical context of our corpus linguistic approach to mind-modelling (Mahlberg & Stockwell 2016; Stockwell & Mahlberg 2015). In a comparison of Dickens's novels against a corpus of other 19th century novels (19C), the 19C corpus is used to model the background knowledge about fictional characters in 19th century fiction that readers might bring to the experience of reading Dickens.

This case study illustrates the application of the method to a concrete research question, the purpose being not to develop the theoretical arguments further but to make the case for the usefulness of the method in a practical context. While we use a case study of literary texts for illustration in this paper we will also indicate how our approach to collocation has implications for the conceptualisation of discourse more widely.

References

Baroni, M., & Bernardini, S. (2003). A preliminary analysis of collocational differences in monolingual comparable corpora. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), Proceedings of the Corpus Linguistics 2003 Conference. Lancaster: UCREL, Lancaster University, 82-91.

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. Journal of English Linguistics, 36 (1), 5-38.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., … Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1 (1), 7-36.

Mahlberg, M., & Stockwell, P. (2016). Point and CLiC: teaching literature with corpus stylistic tools. In M. Burke, O. Fialho, & S. Zyngier (Eds.), Scientific Approaches to Literature in Learning Environments. Amsterdam: John Benjamins, 251-267.

Stockwell, P., & Mahlberg, M. (2015). Mind-modelling with corpus stylistics in David Copperfield. Language and Literature, 24 (2), 129-147.

• • •

Guyanne Wilson
Ruhr-Universität Bochum
Guyanne.Wilson@ruhr-uni-bochum.de

"There's different types": Agreement in *there* + *be* existential constructions in Nigerian English.

English existential constructions typically comprise a syntactic subject followed by a form of the verb *to be* and a notional subject, as in

There	*is*	*a hole in the bucket.*
syntactic subject	V	notional subject

In such constructions, subject-verb agreement is thought to be governed by the notional subject. Studies on different inner-circle Englishes, however, have shown that existential constructions often exhibit non-agreement, particularly in spoken language, where constructions like, *"There's holes in the bucket,"* are not uncommon (cf. Meecham and Foley 1994 (Ottowa), Hay and Schreier 2004 (New Zealand) Cheshire and Fox 2009 (London)), so that some researchers have labelled the use of a singular verb preceding a plural noun phrase in existential constructions a "vernacular universal" (Walker 2007: 163).

Among outer-circle Englishes, existential non-agreement has been found to be less frequent than in inner-circle varieties (cf. Jantos 2009, Collins 2012). At the same time, previous studies have focused mostly on Asian and SE Asian Englishes, notably excluding African Englishes. The publication of the International Corpus of English Nigerian corpus (ICE Nigeria) means that this variety may now be included in general analysis.

This paper presents the results of a study looking at agreement in existential constructions in Nigerian English. It compares the spoken and written components of ICE Nigeria and looks at the overall frequency of non-agreement in existential constructions as well as differences between the two registers, before looking more closely at agreement across four ICE text categories: S1A (private dialogues), S1B (public dialogues), unscripted monologues (S2A), and scripted monologues (S2B).

Findings show overall rates of non-agreement in Nigerian English to be 2.68% in written texts and 4.39% in spoken texts, the latter lower than the 12 and 17 percent Jantos (2009) reported for British and American English respectively, though within the range reported for outer-circle Englishes when only S1A is considered. Cross-categorical comparisons reveal patterning similar to that found in ICE Singapore: in less formal, more spontaneous text categories (S1A and S2A), rates of non-agreement are higher than in more formal text categories with a higher degree of planning. Like other outer-circle varieties, Nigerian English further exhibits non-agreement in plural existential constructions, i.e. constructions such as *There are a hole in the bucket*,

reported as restricted to older, rural speakers, and to past tense constructions in inner-circle Englishes (e.g. Durham 2013, Tagliamonte 1998).

That Nigerian English behaves more like outer-circle than inner-circle Englishes allows us to interrogate the notion of "angloversals" (Mair 2003) more closely. To this end, different constraints acting on non-agreement— plurality, contractedness, and tense— will also be examined, to see whether they are the same as in other varieties. Additionally, rates of non-agreement in other types of constructions will be inspected, in order to ascertain whether the non-agreement, even in comparatively low frequencies, is a feature peculiar to existential constructions, or is instead a general feature of Nigerian English.

References

Cheshire, J., & Fox, S. (2009). Was/were variation: a perspective from London. Language Variation and Change 21(1): 1-38.

Collins, P. (2012). Singular agreement in there-existentials: an intervarietal corpus-based study. English World-Wide 33(1), 53–68.

Durham, M. (2013). Was/were alternation in Shetland English. World Englishes 32(1), 108-128.

Hay, J., & Schreier, D. (2004). Reversing the trajectory of language change: Subject–verb agreement with be in New Zealand English. Language variation and change 16(3), 209-235.

Jantos, S. (2009). Agreement in educated Jamaican English: a corpus investigation of ICE Jamaica. PhD Dissertation: Albert-Ludwigs-Universität Freiburg. Available online at https://freidok.uni-freiburg.de/dnb/download/7588.

Mair, C. (2003). Kreolismen und verbales Identitätsmanagement im geschriebenen jamaikanischen Englisch. In E. Vogel, A. Napp & W. Lutterer (Eds.), Zwischen Ausgrenzung und Hybridisierung, 79-96. Ergon.

Meechan, M. & Foley, M. (1994). On resolving disagreement: linguistic theory and variation–There's bridges. Language variation and change 6(3), 63-85.

Tagliamonte, S. (1998). Was/were variation across the generations: view from the city of York. Language variation and change, 10(2), 153–191.

Walker, J. A. (2007). „There's bears back there": Plural existentials and vernacular universals in (Quebec) English. English World-Wide 28(2), 147-166.

• • •

Work-in-progress

Karin Axelsson
University of Gothenburg
karin.axelsson@sprak.gu.se

Tag question in Spoken BNC2014 Early Access Subset

The *British National Corpus* (BNC) and the British component of the *International Corpus of English* (ICE-GB) have been utilised several times to study the use of canonical tag questions (TQs) in British English conversation. The spoken demographic part of BNC (henceforth BNC1994D) is used by e.g. Tottie and Hoffmann (2006) and Axelsson (2011), and spoken parts of ICE-GB by e.g. Gómez-González (2012) and Kimps (2016). The contraction *innit* (derived from *isn't it/ain't it* and therefore also regarded as canonical here) has received a great deal of attention during the last decades, based on data from BNC, the *Bergen Corpus of London Teenage Language* (COLT) and the *Linguistic Innovators Corpus* (LIC) (e.g. Krug 1998, Andersen 2001, Torgersen *et al.* 2011, Pichler 2013, Palacios Martínez 2015). Due to the development of *innit* and as BNC, COLT and ICE-GB reflect English two decades ago and as LIC is restricted to the London area and focuses mainly on teenagers, it is high time to study canonical TQs including *innit* in contemporary conversation throughout Britain.

The study of TQs presented here is based on data from a 5-million-token Early Access Subset of a new corpus of British English conversation: *Spoken BNC2014* (Love *et al.* 2017 fc.). After random thinning, 497 TQs with declarative anchors, DecTQs, (including *innit*) were identified and analysed. As some speakers with very many instances in the dataset skewed the results severely, all speakers with more than 75,000 words were removed, and a subcorpus called BNC2014ER with 362 speakers was created: this reduced dataset contains 238 TQs. The normalised frequency of DecTQs (including *innit*) in BNC2014ER is 2,747 per million words (pmw), i.e. much lower than in BNC1994D: 5,062 pmw. As there were so few instances of *innit* in this dataset, a separate search for all instances in BNC2014R was conducted, resulting in a frequency of just 76 pmw (N=203) compared to 396 pmw (N=114) in BNC1994D. These low frequencies indicate that both *innit* and other DecTQs are losing ground. However, one may question whether the two BNC corpora are completely comparable for the study of DecTQs since they have been compiled in different ways. BNC2014E uses crowd-sourcing where the requirements of good recording quality seem to favour more focused conversations than in BNC1994D, where the respondents were chosen randomly and told to record all spoken interactions during two days. The low number of *innit* might partly be due to transcription guidelines saying "only use *innit* when you are sure: otherwise either use *isn't it* or *ain't it*" in combination with the clearer recordings.

Apart from mere frequencies, the present study also compares formal features, functions and sociolinguistics factors. DecTQs in general are used to similar extents by both genders, whereas men use *innit* significantly more often than women. Tottie and Hoffmann (2006:304) found that, in BNC1994D, speakers older than about 25 used more TQs than younger speakers. The data in BNC2014ER indicates that the cut-off

between younger and older speakers is now instead around the age of 40. *Innit* is most frequent in the age group 19–29 and then gradually decreases. There are no significant differences as to social grade for DecTQs in general, whereas *innit* is associated with the lower grades.

References

Andersen, G. (2001). Pragmatic Markers and Sociolinguistic Variation: a Relevance-Theoretic Approach to the Language of Adolescents. Amsterdam: Benjamins.

Axelsson, K. (2011). Tag Questions in Fiction Dialogue. PhD, University of Gothenburg, Göteborg. URL: http://hdl.handle.net/2077/24047.

Gómez-González, M. (2012). The question of tag questions in English and Spanish. In I. Moskowich & B. Crespo (Eds.), Encoding the Past, Decoding the Future: Corpora in the 21st Century. Newcastle upon Tyne: Cambridge Scholarly Press, 59–97.

Kimps, D. (2016). English variable tag questions: a typology of their interpersonal meanings. PhD, KU Leuven, Leuven.

Krug, M. (1998). British English is developing a new discourse marker, innit? A study in lexicalisation based on social, regional and stylistic variation. Arbeiten aus Anglistik und Amerikanistik, 23 (2), 145–197.

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

Palacios Martínez, I. M. (2015). Variation, development and pragmatic uses of innit in the language of British adults and teenagers. English Language and Linguistics, 19 (3), 383–405.

Pichler, H. (2013). The Structure of Discourse-Pragmatic Variation. Amsterdam: Benjamins.

Torgersen, E. N., Gabrielatos, C., Hoffmann, S., & Fox, S. (2011). A corpus-based study of pragmatic markers in London English. Corpus Linguistics and Linguistic Theory, 7 (1), 93–118.

Tottie, G., & Hoffmann, S. (2006). Tag questions in British and American English. Journal of English Linguistics, 34 (4), 283–311.

• • •

Cinzia Bevitori
University of Bologna
cinzia.bevitori@unibo.it

Exploring ‚justice' in parliamentary debates on im/migration through the lens of corpora

The paper is a by-product of a new interdisciplinary research project aiming at critically assessing the European Union's impact on global justice by looking at different areas and bringing together various disciplines and methodologies. One of such areas is ‚migration' (broadly including asylum and refugees), an issue which has been raising a number of dilemmas for the EU and its member states concerning justice as it hints at fundamental and possibly conflicting principles within the Union. Although a number of corpus-based studies, from critical discursive analytical perspectives, have focused on representations of migrants and asylum seekers, they have mostly concentrated on newspaper discourse (e.g. Baker et al 2008, Taylor 2014), as well as on discourses of discrimination in different institutional contexts (e.g. Wodak and van Dijk eds. 2000, Kryzanowski and Wodak 2008). Moving from the assumption that justice is a 'human construction' (Waltzer 1986), and that different theoretical conceptions of justice are at stake (Eriksen 2016), the paper will attempt to examine discourses surrounding the issue of migration in a wider context of 'global justice' looking at data from a corpus of UK parliamentary debates consisting of the complete transcriptions of all the debates held in the House of Commons on the subject of 'im/migration' in 2015 - a key year for the migration and refugees crisis in Europe. The analysis combines quantitative and qualitative dimensions of investigation, by drawing on the tools and techniques of corpus-assisted discourse analysis (inter alia, Baker 2006, Baker et al 2008, Baker and McEnery 2015, Morley and Bayley eds 2009, Partington, Morley and Haarman eds 2004, Thompson and Hunston eds 2006). The aim of the study is two-fold. In the first place, it will discuss what (and how) understandings of justice emerge from the discourses regarding 'im/migrants' in this particular institutional and political sub-domain of political discourse. Parliamentary discourse is, in fact, a very distinctive and composite type of political discourse regulated by long-standing conventions (Bayley ed. 2004), which may be rightly considered as a privileged site of analysis of the 'struggle over meanings' (Miller 1997; see also Bevitori 2005, 2006, 2007). In the second place, it will also reflect on the methodological challenge of analysing complex discursive issues through a corpus-assisted approach.

References

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

Baker, P. & McEnery T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. Journal of Language and Politics, 4(2): 197-226.

Baker, P. & McEnery, T. (2015). Corpora and Discourse. Integrating Discourse and Corpora. Palgrave MacMillan UK.
Baker, P., Gabrielatos, C. , Khosravinik, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (2008). A useful synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse and Society, 19(3), 273-306.
Bayley, P. (2004). Cross-cultural Perspectives on Parliamentary Discourse. Amsterdam: John Benjamins.
Bevitori, C. (2005). Attribution as evaluation: a corpus-based investigation of quotations in parliamentary discourse. ESP Across Cultures, 2, 7–20.
Eriksen, E.O. (2016). Three Conceptions of Global Political Justice. ARENA Centre for European Studies, University of Oslo. GLOBUS Research Paper 1/2016.
Krzyżanowski, M &. Wodak, R. (2009). The Politics of Exclusion: Debating Migration in Austria. New Brunswick, NJ: Transaction Publishers.
Hoey, M., Mahlberg, M., Stubbs, M. & Teubert W., (2007). Text, Discourse and Corpora. Theory and Analysis. London: Continuum.
Miller D.R. (1999). Meaning up for grabs: Value-orientation patterns in British parliamentary debates on Europe. In J. Verschueren (ed.), Language and Ideology: Selected Papers from the 6th International Pragmatic Conference. Vol. 1, 386-404. IPrA: Antwerp, Belgium.
Miller, D. R., Bayley, P., Bevitori, C., Fusari, S. & Luporini, A. (2014). 'Ticklish trawling': The limits of corpus assisted meaning analysis'. In S. Alsop & S. Gardner (eds), Language in a Digital Age: Be Not Afraid of Digitality: Proceedings from the 24th European Systemic Functional Linguistics Conference and Workshop, Coventry University: Coventry, UK. https://curve.coventry.ac.uk/open/items/7b5b94aa-6984-48ad-b29a-9a8e9483fa2d/1/ ISBN: 978 18460007 13.
Morley, J. & Bayley P. (eds) (2009). Corpus-assisted discourse studies on the Iraq conflict: Wording the war. New York: Routledge.
Partington, A., Morley, J. & Haarrman, L. (eds) (2004). Corpora and Discourse. Bern: Peter Lang.
Stubbs, M. (1996). Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture. Oxford: Blackwell Publishing.
Taylor, C. (2014). Investigating the representation of migrants in the UK and Italian press: A cross-linguistic corpus-assisted discourse analysis'. International Journal of Corpus Linguistics, 19(3), 368-400.
Thompson, G. & Hunston, S. (2006). System and Corpus: two traditions with a common ground. In G. Thompson & S. Hunston (eds.), System and Corpus: exploring connections. London: Equinox, 1-14.
Walzer, M. (1983). Spheres of justice: A defence of pluralism and equality. Oxford: Robertson.
Wodak, R. & van Dijk, T. (eds.) (2000). Racism at the Top: Parliamentary Discourses on Ethnic Issues in Six European States. Drava Verlag: Klagenfurt.

• • •

Samuel Bourgeois
University of Neuchâtel
samuel.bourgeois@unine.ch

How discourse markers cross into writing: Colloquialization and the development of *actually*

This study investigates how the discourse marker (DM) *actually* is increasingly adopted into written genres and what functional changes go along with this development. Previous work on the multifunctionality of *actually* has discussed how it diachronically develops from adverbial senses to "epistemic adversative" senses and further to DM "additive" senses (Traugott and Dasher 2002: 169-170). In synchrony, DM *actually* has been analyzed primarily in studies focusing on conversational data (Tognini-Bonelli 1993; Smith and Jucker 2000; Clift 2001), which is motivated by the close association that DMs have with oral genres (Schourup 1999: 234). Despite the obvious association of DMs with orality, however, Aijmer (2013) argues that DMs have a meaning potential that allows their adoption into new genres and new functions. This paper will expand on this idea and it will argue on the basis of corpus data that the use of DMs in writing is an example of colloquialization (Mair 2006: 186).

Methodologically, this study adopts the outlook of corpus pragmatics (Rühlemann and Aijmer 2015), but confines this approach to the less visited arena of written texts. In particular, it will concentrate on the uses of *actually* in the written sections of the COHA (Davies 2010). The first part will compare the data from the COHA and the Hansard Corpus (Alexander and Davies 2015), which allows one to observe *acutally*'s behavior diachronically from the 19th century to the first decade of the 21st century in both written and oral form. Through this comparison, it will also highlight the particularities of *actually* as used in writing. Furthermore, it is here that it will be demonstrated that in writing particularly there is a general increase in the frequency of use of *actually* since the second half of the 20th century. The second part of this study will take a qualitative approach and will present an in-depth look into how *actually* functions in writing in the sentence initial, medial, and final positions. Special attention will be given to *actually* in the sentence medial position because previous studies of *actually* have demonstrated that its use in the medial position in writing differs from its use in conversation (Oh 2000; Kallen 2015).

Ultimately the finding that emerges from this analysis is that colloquialization is more than the inclusion of oral elements into writing. As DMs like *actually* make their way into written genres, their functions adapt to the specific communicative needs of writers. Particularly this work will highlight the marked rise of importance of *actually* being used to mark clause boundaries and word selection, a practice also demonstrated to be increasingly frequent with the DM *well* since the later 20th century (Rühlemann and Hilpert *to appear*). Furthermore, the data demonstrates that sentence medial *actually* serves the double function of mimicking the conversational-specific functions that have to do with upgrading or correcting terminology used by speak-

ers, while also serving the uniquely written function of signaling a salient syntactical boundary or intentional lexical selection.

References

Aijmer, K. (2013). Understanding pragmatic markers. A Variational Pragmatic Approach. Edinburgh: Edinburgh University Press.

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

Clift, R. (2001). Meaning in Interaction: The case of actually. Language 77(2), 245-91.

Davies, M. (2010). The Corpus of Historical American English (COHA) 1810-2009. Available online at http://corpus.byu.edu/coha.

Kallen, J. (2015). 'Actually, it's unfair to say that I was throwing stones': Comparing Perspectives on the uses of actually in ICE-Ireland. In C. Amador-Moreno, K. McCafferty & E. Vaughan (Eds.), Pragmatic Markers in Irish English. Amsterdam and Philadelphia: John Benjamins Publishing Company, 135-55.

Mair, C. (2006). Twentieth-Century English: History, Variation, and Standardization. Cambridge: Cambridge University Press.

Oh, S. (2000). Actually and in fact in American English: A data-based analysis. English Language and Linguistics 4(2), 848-78.

Rühlemann, C. & Aijmer K. (2015). Corpus pragmatics: laying the foundations. In K. Aijmer & C. Rühlemann (Eds.), Corpus Pragmatics: A Handbook. Cambridge: Cambridge University Press, 1-26.

Rühlemann, C. & Hilpert, M. (forthcoming). Colloquialization in Journalistic Writing: The Case of Inserts with a focus on well. Journal of Historical Pragmatics.

Schourup, L. (1999). Tutorial Overview: Discourse Markers. Lingua 107, 227-265.

Smith, S. & Jucker, A. (2000). Actually and other markers of an apparent discrepancy between propositional attitudes of conversational partners. In G. Anderson, & T. Fretheim (Eds.), Pragmatic Markers and Propositional Attitude. Amsterdam and Philadelphia: John Benjamins Publishing Company, 207-37.

Tognini-Bonelli, E. (1993). Interpretative nodes in discourse: Actual and actually. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), Text and Technology: In Honour of John Sinclair. Amsterdam and Philadelphia: John Benjamins Publishing Company, 193-211.

Traugott, E. & Dasher, R. (2001). Regularity in Semantic Change. Cambridge: Cambridge University Press.

• • •

Elena Callegaro
University of Zurich
elena.callegaro@es.uzh.ch

Simon Clematide
University of Zurich
siclemat@ifi.uzh.ch

The validity of large data-driven and constructional approaches for the investigation of variable article use in English

Articles are among the most frequently used words in the English language (OED Online 2016). Standard grammars agree on the fact that the definite article is used in front of NPs to express definiteness, whereas the indefinite article expresses indefiniteness. The zero article, on the other hand, gives a conception of a whole class with a general connotation (Quirk et al. 1985; Biber et al. 1999; Huddleston and Pullum 2002). While this distinction seems fairly evident, articles have received considerable critical attention due to their complexity in usage; however, there is still a general lack of research on their (variable) use (but see Tse 2001, 2003, 2004; Yoo 2007; Hundt 2016, in press; Callegaro et al. forthcoming).

This paper builds on automatically annotated and aligned data from *Costep* (Graën, Batinic and Volk 2014), an improved version of the large parallel *Europarl* corpus (Koehn 2005). *Costep*'s refined information about the speakers' nationality allows for the distinction between original texts and translated material. In a current study (Callegaro, in preparation), *Costep* enables us to retrieve bare cases, which are notoriously hard to find in English, by using cross-language evidence from German, a language that uses articles frequently (Duden Online 2016). Using German as a starting point permits us to retrieve the contexts in which an article is used and their corresponding aligned parallel instances in English in which an article does not occur. In the present study, a combination of a similar data-driven approach with a Construction Grammar perspective is applied to detect highly variable article use in present-day English. This corpus-linguistic procedure results in a list of head nouns that occur almost equally frequently with or without an article. Based on a qualitative selection, we then narrow down our analysis on variable article use with abstract non-count nouns (e.g. *regeneration, education*) followed (or not) by the preposition *of* (e.g. *attention* vs. *the attention of*). As also described by Quirk et al. (1985: 286-7), these nouns normally occur as bare NPs, but require an article when postmodified by an *of*-phrase construction (e.g. *She's studying European history* vs. *She's studying the history of Europe* and not **history of Europe*). The largely data-driven method is thus useful because it "aims to derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context" (Tognini-Bonelli 2001: 87). Furthermore, the constructional approach sheds new light on the topic of language variation (and in particular variable article use), which "has only recently been put on the research agenda of Construction Grammarians, who

are thus relative late-comers" (Hilpert 2014: 185). As Trousdale and Gisborne (2008: 71) state, "corpora can provide empirical support to intuitions regarding the nature of constructions and the number of constructions in the constructional inventory." Therefore, bringing together these approaches for the investigation of variable article use can be considered a significant empirical contribution to this growing research area.

The present study is thus an assessment of the validity of this empirical method for the analysis of linguistic variation. Moreover, this investigation aims to advance the understanding of variable article use in English from a Construction Grammar point of view, using data from an innovative corpus. Finally, the cross-language setup of this parallel corpus enables us to contrast and compare our findings for English constructions with other languages.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan E. (1999). The Longman Grammar of Spoken and Written English. London: Longman.

Callegaro, E., Clematide, S., Hundt, M., & Wick, S. (forthcoming). Variable article use with acronyms and initialisms – a contrastive analysis of English, German and Italian. Languages in contrast.

Callegaro, E. (In preparation). Variable Article Use – A contrastive Study of English and German. PhD dissertation, University of Zurich.

Duden Online. (2016). Die häufigsten Wörter in deutschsprachigen Texten. <http://www.duden.de/sprachwissen/sprachratgeber/die-haeufigsten-woerter-in-deutschsprachigen-texten> (accessed 9 February 2016)

Hilpert, M. (2014). Construction grammar and its application to English. Edinburgh: Edinburgh University Press.

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

Hundt, M. (2016). Who is the/a/ø professor at your university? A construction-grammar view on changing article use with single role predicates in American English. In M. J. López-Couso, B. Méndez-Naya, P. Núñez-Pertejo & I. M. Palacios-Martínez, (Eds.), Corpus Linguistics on the Move: Exploring and Understanding English Through Corpora. Amsterdam & New York: Brill/Rodopi, 227–258.

Hundt, M. (In press). Variable article usage with institutional nouns – an 'oddment' of English?

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In MT summit. Vol. 5. 79–86.

OED Online. (2016). The OEC: Facts about the language. Oxford University Press <http://www.oxforddictionaries.com/words/the-oec-facts-about-the-language> (accessed 9 February 2016)

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., & Crystal, D. (1985). A comprehensive grammar of the English language. Vol. 397. London: Longman.

Tognini-Bonelli, E. (2001). Corpus linguistics at work. Vol. 6. Amsterdam: Benjamins.

Trousdale, G., & Gisborne, N. (Eds.). (2008). Constructional approaches to English grammar. Vol. 57. Berlin: de Gruyter.
Tse, G.Y.W. (2001). The grammatical factors influencing the choice between the use and omission of the definite article preceding multi-word organization names: a statistical analysis. Journal of Quantitative Linguistics, 8 (1), 13–32.
Tse, G.Y.W. (2003). Validating the logistic model of article usage preceding multi-word organization names with the aid of computer corpora. Literary and Linguistic Computing, 18 (3), 287–313.
Tse, G.Y.W. (2004). A grammatical study of personal names in present-day English: with special reference to the usage of the definite article. English studies, 85 (3), 241–259.
Yoo, I.W.H. (2007). Definite article usage before last/next time in spoken and written American English. International Journal of Corpus Linguistics, 12 (1), 83–105.

• • •

Jarle Ebeling
University of Oslo
jarle.ebeling@usit.uio.no

Bringing closure

This WiP will report on a study of the sequence V + *closure*, where the meaning is achieve „a sense of personal resolution" (OED). It is situated within the framework of empirical phraseology and Sinclair's (e.g. Sinclair 1991; 1996) notion of extended lexical units. In addition, it will draw on insights gained from two recent papers by Stubbs (2013; 2014). The latter, in particular, "compares central concepts in the work of John Searle […] and John Sinclair […] and considers whether these concepts are compatible" (Stubbs 2014: 243).

A search for the word *closure* in the Corpus of Historical American English (COHA) yields a striking result regarding its frequency of use. It has gone from 0 to 5.45 occurrences per million words in approx. 200 years. A sharp increase is especially noticeable between the 1980s and the 1990s (2.53 vs. 4.78 occ. per mill. words).

According to the Oxford English Dictionary (OED), the word is attested as far back as Chaucer and has a wealth of (related) meanings to do with that which encloses or the act of (en)closing or ending. The meaning we are interested in was added in 2006 by the editors of the OED, and reads:

> orig. U.S. A sense of personal resolution; a feeling that an emotionally difficult experience has been conclusively settled or accepted. In early use chiefly Psychoanal.

One example is:

(1) 1987 Social Casewk. 68 546/2 The social worker's goal should be to help bring closure to relationships in such a way as to minimize unresolved feelings and issues.

In his discussion of extended lexical units such as "par for the course" and "smelling of roses", Stubbs (2014) underlines that "if we could not interpret the connotations of such phraseology, we would not be able to understand how other people interpret the social world and would be permanently socially disoriented" (ibid. 256). This, it seems, ties in with Searle's notion of how we create and maintain social "institutions" ("objective cultural reality", Stubbs ibid. 252), e.g. marriage or professorship. We create state of affairs (reality/meaning) when we utter, write and repeat near-equal sequences of words, whether this state of affair is an institution of a more obvious kind or an experience encoded as "coming out smelling of roses" or "bringing closure". "[T]hey are all social constructs maintained by language" (ibid. 253). What Stubbs misses in both Searle's and Sinclair's work is "empirical research – both textual and ethnographic" (ibid. 257), since the units of meaning we are discussing here can be, and most likely are, tied to different categories of speech acts (Searle 1979), can appear in different genres or registers of text and may have a (clear) cultural basis.

Drawing on data extracted from the Corpus of Contemporary American English (COCA), this WiP report will explore V + *closure* as a speech act of the Directive type and as a an emerging extended lexical unit, where it has the meaning quoted from the OED. The focus will be on the type of verbs closure collocates most frequently with and the pattern's frequency and distribution in COCA between 1990 and 2015.

Sources

COCA: Davies, M. (2008-) The Corpus of Contemporary American English: 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

COHA: Davies, M. (2010-) The Corpus of Historical American English: 400 million words, 1810-2009. Available online at http://corpus.byu.edu/coha/.

OED Online. Oxford University Press, March 2017.

References

Searle, J.R. (1979). Expression and meaning. Studies in the Theory of Speech Acts. Cambridge: CUP.

Sinclair, J.McH. (1991). Corpus, Concordance, Collocation. Oxford: OUP.

Sinclair, J.McH. (1996). The search for units of meaning. Textus IX: 75-106.

Stubbs, M. (2013). Sequence and order: The neo-Firthian tradition of corpus semantics. In H. Hasselgård, J. Ebeling & S.O. Ebeling (Eds), Corpus Perspectives on Patterns of Lexis. Amsterdam: Benjamins, 13-34.

Stubbs, M. (2014). Searle and Sinclair on communicative acts. A sketch of a research problem. In M. Gómez González, F. J. R. dM. Ibáñez, F. Gonzálvez García & A. Downing (Eds), The Functional Perspective on Language and Discourse. Applications and Implications. Amsterdam: Benjamins, 243-260.

• • •

Robert Fuchs
Hong Kong Baptist University
robert.fuchs.dd@gmail.com

Alexandra Esimaje
Benson Idahosa University
alexandra.esimaje@live.com

Enhancing corpus data through perception studies – a look at intensifier strength in British and Nigerian English

There is a growing body of evidence showing that patterns of intensifier usage differ between varieties of English, with Biber et al. (1999), for example, reporting differences in intensifier usage between British and American English. Moreover, several studies found that intensifiers occur less frequently in Outer Circle Englishes (with English as a Second Language) than in Inner Circle Englishes (with English as a Native Language): De Klerk (2005) observed that Xhosa speakers of English use a lower number of both types and tokens of intensifiers compared to New Zealand English speakers, while Coronel (2011) showed an overall lower rate but a wide lexical range of intensifiers in spoken Philippine English. Fuchs and Coronel (2011) found lower intensification rates in five Outer Circle Englishes compared to three Inner Circle Englishes, such as between Nigerian English, on the one, and British English, on the other hand.

These and other studies demonstrate regional variation in intensifier usage, but focus exclusively on production. To our knowledge, no previous studies have investigated the perception of intensifier strength, and how it differs

(1) between intensifier types (such as *very, really*),
(2) between varieties of English,
(3) whether age and gender influence the perception of intensifier strength and
(4) whether syntactic position and the identity of the word that is intensified influence perceived intensifier strength.

Studying differences in the perception of intensifiers has the potential to greatly enrich the quantitative perspective afforded by corpus studies with qualitative findings, permitting a reevaluation of the corpus data. More specifically, while Fuchs and Coronel (2011) found a lower intensification rate in Nigerian compared to British English, it is at least conceivable that perceived intensifier strength (which we conceive of as a linear scale from low to high) is a confounding factor offsetting part of this difference. For example, Nigerians might use intensifier types with greater perceived intensification strength relatively more often than Britons.

To answer these questions, this paper reports the results of a perception study on intensifier strength, and its application to corpus data from the Nigerian and British components of the International Corpus of English (Fuchs and Coronel 2011). 102 Nigerians (all proficient speakers of English) and 52 Britons rated the perceived intensification strength of the 40 most frequent boosters and maximisers in the two

corpora on a ten-point scale. Mixed effects regression models were used to determine whether the factors named above influence perceived intensification strength.

Results show that

1. Age, gender, the identity of the intensified word, and syntactic position (attributive or predicative) do not significantly influence intensification rate.
2. The British, but not the Nigerian respondents, rated maximisers (such as *extremely*) as having greater intensification strength than boosters (such as *very*).
3 The Nigerian, but not the British respondents, rated expletives (*bloody*, *fucking*) and low frequency intensifier types as having lower intensification strength than high frequency types.

The application of these results to the corpus data shows that, in NigE, intensifier types with relatively high perceived strength are used comparatively more often than in BrE. This suggests that some of the difference in intensification rate between NigE and BrE observed in the production data might be offset by differences in the perception of intensifier strength. We reinterpret the corpus results in the light of these findings, and conclude that a difference in intensification rate persists between the two varieties, albeit at a lower level.

References

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). Longman Grammar of Spoken and Written English. London: Longman.

Coronel, L. (2011). Patterns of intensifier usage in Philippine English. In Bautista, M. Lourdes S. (ed.), Studies of Philippine English: Exploring the Philippine Component of the International Corpus of English. Manila: De La Salle University Centennial Book Series, 93-116.

Fuchs, R. & Coronel, L. (2011). Intensifier usage across varieties of English. Paper presented at the conference of the International Association for World Englishes (IAWE 2011), Melbourne, Australia.

De Klerk, V. (2005). Expressing levels of intensity in Xhosa English. English World-Wide 26, 77-95.

• • •

Dana Gablasova
Lancaster University
d.gablasova@lancaster.ac.uk

Vaclav Brezina
Lancaster University
v.brezina@lancaster.ac.uk

MI-score-based collocations in language learning research: A critical evaluation

Formulaic language has occupied a prominent role in the study of language learning and use for several decades (Wray, 2013). Recently an even more notable increase in interest in the topic has led to an 'explosion of activity' in the field (Wray 2012: p.23). Language learning research (LLR) in both first and second language acquisition has focused on examining the links between formulaic units and fundamental cognitive processes in language learning and use, such as representation of and access to these units in mental lexicon (Wray 2002, 2012, 2013; Ellis et al. 2015). Collocations, a specific unit of formulaic language, hold a prominent position in LLR, having been used in a number of studies on formulaicity in L2 (Schmitt 2012). The corpus-based measures for identifying collocations (i.e. association measures) in these studies are of paramount importance as they directly and significantly affect the findings of these studies and consequently the insights into language learning that they provide. One of the most prominent and frequently selected association measures in these studies is the Mutual Information score (MI-score), often referred to as a measure of collocational 'strength' (c.f. Hunston 2002).

While the MI-score has been a useful measure in LLR, there are also several issues related to its use (Gablasova et al. forthcoming 2017). First, the rationale behind the selection of the MI-score in studies on formulaic development is not always fully transparent and systematic (González Fernández & Schmitt 2015) and often motivated by tradition rather than by specific aims of a given LLR study. Second, alternative measures are rarely considered and their relevance to LLR is not further examined (Gilquin & Gries 2009). Finally, the application and interpretation of the MI-score in LLR suggests that a fuller understanding of the mathematical and linguistic principles on which the measure is based is needed in LLR studies (e.g. an understanding of what type of collocations receive higher MI values and the reasons for this). This understanding would enable a better interpretation of collocational patterns found in L2 production.

In order to address these issues, the paper seeks to achieve the following objectives: i) to place the MI-score in the context of other similar association measures and discuss the similarities and differences directly relevant to LLR; ii) to propose general principles for selection of association measures in LLR. The study examines these questions using data from several corpora and sub-corpora (e.g. the BNC and the Trinity Lancaster Corpus of L2 spoken English). Using these corpora, we examine the

linguistic patterns identified by the MI-score and contrast them with other association measures (e.g. Log Dice) paying special attention to how collocational patterns (i.e. collocational strength) changes according to different measures used.

References

Ellis, N.C., Simpson-Vlach, R., Römer, U., Brook O'Donnell, M. & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition. In S. Granger, G. Gilquin & F. Meunier (Eds.), The Cambridge Handbook of Learner Corpus Research (pp 357-378). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.016

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: identifying, comparing and interpreting the evidence. Language Learning.

Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. (2015). Epistemic stance in spoken L2 English: The effect of task type and speaker style. Applied Linguistics (Advance Access). doi:10.1093/applin/amv055

Gilquin, G., & Gries, S. Th. (2009). Corpora and experimental methods: A state-of-the-art review. Corpus Linguistics and Linguistic Theory, 5(1), 1-26. doi:10.1515/CLLT.2009.001

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have?: The effects of frequency and amount of exposure. International Journal of Applied Linguistics, 166(1), 94-126. doi:10.1075/itl.166.1.03fer

Hunston, S. (2002). Corpora in applied linguistics. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524773

Schmitt, N. (2012). Formulaic Language and Collocation. In C. Chapelle (Ed.), The Encyclopedia of Applied Linguistics. New York: Blackwell. doi:10.1002/9781405198431.wbeal0433

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

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. Annual Review of Applied Linguistics, 32, 231-254. doi:10.1017/S026719051200013X

Wray, A. (2013). Formulaic language. Language Teaching, 46(3), 316-334. doi:10.1017/S0261444813000013

• • •

Gunther Kaltenböck
University of Vienna
gunther.kaltenboeck@univie.ac.at

***Funny that you should say that*: On the use of semi-insubordinate clauses**

Insubordinate clauses have received increased interest in recent years (e.g. Evans 2007, Mithun 2008, D'Hertefelt & Verstraete 2014) with the term being applied to a wide range of different constructions including so-called semi-insubordinate constructions as in (1) (Van linden & Van de Velde 2014, Sansiñena 2015), which have received little attention in English.

(1) a. Well, funny you should ask, Florence. (COCA:1999:SPOK:NBCToday)
b. Strange that Lucinda should not want to stay in London (BNC:CEH W fict)
c. Shame Tom wasn't in (BNC:KC7 S conv)

The aim of the present paper is two-fold: (i) to investigate the formal and functional properties of these constructions and (ii) to critically examine whether they qualify for inclusion into the category of insubordination, as has been suggested by Van linden & Van de Velde (2014). The database for the study is provided by a range of different corpora, notably the British National Corpus, the Corpus of Contemporary American English and the Corpus of Historical American English.

It is possible to distinguish a number of different formal subtypes based on the syntactic category of the initial predicate (adjective, noun) and the type of complement clause (that, zero, how, to-infinitive, -ing clause), which in turn relates to different uses of the construction. Functionally, the construction is shown to be used in different ways with the prototypical function being that of a subjectivising presentative construction, where the complement clause conveys new information (often introducing a new discourse topic) which is anchored in subjective speaker perspective expressed by a prospective matrix predicate. Depending on the syntactic and/or prosodic form of the complement clause the new information can be presented either as new or known (presupposed).

In terms of grammatical modelling it is argued that the construction does not fulfil the structural criterion of syntactic independence and is there best treated not as instance of insubordination but as an ellipted it-extraposition (or adjective complement construction) with which it also shares a number of structural and semantic properties. In terms of its discourse function, however, the construction resembles insubordinate clauses in being similarly speaker-centred and subjectivising, which can be captured by their analysis as 'theticals' (Heine et al. 2013). It is also suggested that the best way to account for the close functional and formal links with related structures is in terms of a constructionalist framework (e.g. Goldberg 2006).

References

D'Hertefelt, S. & Verstraete, J.-C. (2014). Independent complement constructions in Swedish and Danish: Insubordination or dependency shift? Journal of Pragmatics 60, 89-102.

Evans, N. (2007). Insubordination and its uses. In I. Nicolaeva (Ed.), Finiteness: Theoretical and Empirical Foundations. Oxford: Oxford University Press, 366-431.

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

Heine, B., Kaltenböck, G., Kuteva, T. & Long, H. (2013). An outline of discourse grammar. In S. Bischoff & C. Jany (Eds.), Functional Approaches to Language. Berlin: Mouton de Gruyter, 175-233.

Mithun, M. (2008). The extension of dependency beyond the sentence. Language 84(1), 69-119.

Sansiñena, M. S. (2015). The multiple functional load of que. An interactional approach to insubordinate complement clauses in Spanish. Doctoral dissertation, KU Leuven.

Sansiñena, M. S., De Smet, H. & Cornille, B. (2015). Between subordinate and insubordinate. Paths toward complementizer-initial main clauses. Journal of Pragmatics 77, 3-19.

Van Linden, A. & Van de Velde, F. (2014). (Semi-)autonomous subordination in Dutch: Structures and semantic-pragmatic values. Journal of Pragmatics 60, 226-250.

Verstraete, J.-C., D'Hertefelt, S. & Van Linden, A. (2012). A typology of complement insubordination in Dutch. Studies in Language 36, 123-153.

• • •

Alexander Kautzsch
University of Regensburg
alexander.kautzsch@ur.de

Introducing CNamON: A Corpus of Namibian Online Newspapers

The linguistic situation in present-day Namibia in Southern Africa is characterized by the co-existence of a multitude of languages, i.e. English, Afrikaans, German, and about twenty-five indigenous African languages of the Niger-Bantu and Khoesan families. Despite this linguistic diversity and although Namibia was never under direct British rule, English was introduced as only official language with independence (*The Constitution of the Republic of Namibia* 1990: Art. 3). Interestingly, the study of English as spoken in Namibia is still in its infancy. Besides a substantial body of research on educational policy (e.g. Frydman 2011; Harlech-Jones 1995; Pütz 1995;

Tötemeyer 2010), the structural peculiarities of English in Namibia have only just begun to be examined: Buschfeld and Kautzsch (2014) investigate language attitudes and use and provide a tentative list of lexical, (morpho-) syntactic and phonological features, Stell (e.g. 2014) examines code-switching, and Kautzsch and Schröder (2016) deliver a first take on variable realizations of short vowels based on ethnicity.

What is still absent, however, is a systematic account of the structural properties of English as used in Namibia. To provide a basis for such an inquiry, the present paper introduces a Corpus of Namibian Online Newspapers (CNamON), which encompasses the content of seventeen news sources as available on the Internet from May to June 2016. These sources add up to a corpus of about 44 million words of text. Since these data represent with written, edited language, any peculiarities found cannot easily be dismissed as learner mistakes (cf. Gries and Bernaisch [2016: 7] on a South Asian newspaper corpus) but might be regarded as potential features of Namibian English.

In this presentation, the focus is twofold. The first part addresses the technical details of the compilation process as well as the set-up of the corpus. To demonstrate the potential of this corpus, part two provides a first qualitative stock-taking as well as exemplary quantitative analyses of structural characteristics of English in Namibia on the linguistic levels of lexis, (morpho-) syntax, and discourse.

References

Buschfeld, S. & Kautzsch, A. (2014). English in Namibia: A first approach. English World-Wide, 33(2), 121 160.

Frydman, J. (2011). A critical analysis of Namibia's English-only language policy. In E. G. Bokamba, R. K. Shosted & B. Tesfaw Ayalew, (eds.), Selected Proceedings of the 40th Annual Conference on African Linguistics: African Languages and Linguistics Today. Somerville MA: Cascadilla Proceedings Project, 178-189. (http://www.lingref.com/cpp/acal/40/paper2574.pdf) (2016-10-15).

Gries, S. T. & Bernaisch, T. (2016). Exploring epicentres empirically: focus on South Asian Englishes. English World-Wide, 37(1), 1–25.

Harlech-Jones, B. (1995). Language Policy and Language Planning in Namibia. In M. Pütz (ed.), 181–206.

Kautzsch, A. & Schröder, A. (2016). English in multilingual and multiethnic Namibia: Some evidence on language attitudes and on the pronunciation of vowels. In C. Ehland, I. Mindt & M. Toennies. (eds.), Anglistentag 2015 Paderborn, Proceedings. WVT, Wiss. Verlag Trier, 277 288.

Pütz M. (ed.) (1995). Discrimination Through Language in Africa? Perspectives on the Namibian Experience. Berlin: Mouton De Gruyter.

Stell, G. (2014). Social identities in post-Apartheid intergroup communication patterns: linguistic evidence of an emergent nonwhite pan-ethnicity in Namibia? International Journal of the Sociology of Language 230, 91 114.

The Constitution of the Republic of Namibia. 1990 [2010]. WIPO (World Intellectual Property Organization) Resources. (http://www.wipo.int/wipolex/en/details.jsp?id=9404) (2016-12-01).
Tötemeyer, A.-J. (2010). Multilingualism and the Language Policy for Namibian Schools. (PRAESA Occasional Papers No. 37.). (http://www.praesa.org.za/files/2012/07/Paper37.pdf) (2016-10-10).

• • •

Andrew Kehoe
Birmingham City University
andrew.kehoe@bcu.ac.uk

Matt Gee
Birmingham City University
matt.gee@bcu.ac.uk

Reference and identity in online reader comments: a corpus-based study

In this paper we analyse reference and identity markers in reader comments appearing at the bottom of articles on the website of the UK newspaper *The Guardian* (http://www.theguardian.com/). This is a public forum with over 9 million unique visitors worldwide on a daily basis yet, as we demonstrate, it is an online space where people are able to build distinct sub-communities and get to know one another on an individual basis without necessarily knowing the true identity of the person with whom they are interacting.

Our research is based on a corpus covering the period 2007 to 2010, which includes all 6.2 million comments made on over 500,000 articles published on *The Guardian* website during that time. In a previous study we found that although less than half of *Guardian* articles allow commenting, 85% of those articles have at least one comment and in some sections, such as Sport, more than half of articles have over 40 comments.

We explore in depth the behaviour of individual commenters and the frequency of interaction between them. We begin with an analysis of top commenters, the key finding of which is that, although there are over 470,000 people actively commenting from 2007-2010, 120 of these people are responsible for 10% of all 6.2 million comments in the corpus and 1000 people are responsible for a third of all comments. Particular individuals are extremely active with, for example, the user 'MartynInEurope' making 15,233 comments on 4874 articles. Five further users are responsible for more than 10,000 comments each, with many others in the thousands.

We go on to show how there appear to be distinct sub-communities forming around specific sections of *The Guardian*, each with their own regular contributors who are well known to one another by their chosen usernames. For instance, the user 'CunningStunt' makes 11,500 comments overall, but 98% of these are in the 'Chatterbox' video games section. As we illustrate, members of this sub-community are frequently referred to by username, e.g. in (1) where CunningStunt has advised LazyBones during a previous discussion:

(1) LazyBones: Morning all. Played Chrono-Trigger a bit last night, and that's about it. I laid Cyrus's tormented ghost to rest (thanks **CunningStunt**) and then tried to find the ‚sun stone'.

We carry out a collocational analysis of the top usernames to examine the contexts in which they appear. We find many examples of thanking, as in (1), but also of congratulating, empathising, disagreeing, etc. We then extend our analysis to the articles themselves (using a separate *Guardian* corpus) and find an increasing trend for well-known commenters to be mentioned here too. For instance, (2) comes from a weekly 'best of' article by a *Guardian* journalist summarising recent discussions between Chatterbox commenters.

(2) One of our regulars - who shall remain nameless - asked the ‚box for some advice: „Where is a good place to meet women other than at work or when totally p***ed?" Whilst Dear Deidre isn't in any trouble yet, we had a few sensible suggestions. Salsa or dance classes (thank you **Cunningstunt**), friends of friends (well done **cameroon95**) and **dizzyisanegg's** suggestion of accepting any social invitation, whatever it is, were all good.

We make an important contribution to the growing field of corpus pragmatics by carrying out the first large-scale corpus-based analysis of interactions in reader comments. In doing so, we demonstrate how people commenting under anonymous usernames are able to build rapport and interact regularly, often on a daily basis.

• • •

Aleš Klégr
Filozofická fakulta Univerzity Karlovy
ales.klegr@ff.cuni.cz

Language is embiggened by words that don't exist. The case of a circumfix

The title is borrowed from a short reflection by Laurie Bauer in which he considers the status of words such as *embiggen* which are used by some speakers in some types of English but not recognized by reference books. The circumfix is occasionally mentioned in works on derivational morphology but never exemplified by more than one or two words. Using corpus and web data collected by queries and experiment the paper examines the actual use (and productivity) of the circumfix *en/em*-(Adj)-*en*. Circumfixation by *en/em*-(Adj)-*en* is of interest as a potential means of extending the range of possibilities in which new verbs can be created in contemporary English. It is sometimes claimed that "Virtually the only other means of creating new verbs in English – besides affixation of *-ize* and *-ify* – is conversion" (Lieber, 2004: 89). Other authors add compounding and back-formation, but circumfixation is never mentioned as a candidate. The paper finds that although the circumfix *en/em*-(Adj)-*en* is largely ignored by word-formation specialists (a case in point is Bauer, Lieber and Plag, 2015), it enjoys something of a secret life on the Internet. It is creatively employed for ad hoc purposes on various occasions by native speakers who seem to have specific intuitions about its meaning and use. The paper attempts to formulate the rules (formula) for its use on the basis of a sample of cicumfixed formations collected with the help of a frequency list of English adjectives.

References

Bauer L. (2003). Introducing Linguistic Morphology. Edinburgh: Edinburgh University Press.

Bauer, L., Lieber, R. & Plag, I. (2015). The Oxford Reference Guide to English Morphology. Oxford: Oxford University Press.

Čermák, F. (2008). Diskrétní jednotky v jazyce: případ cirkumfixů. Slovo a slovesnost 69, 1-2, 78-98.

Čermák, F. (2011). Morfematika a slovotvorba češtiny. Nakladatelství lidové noviny, Praha.

Leech, G. (1990). Semantics. Harmondsworth: Penguin Books.

Leech, G., Rayson, P. & Wilson, A. (2001). Word Frequencies in Written and Spoken English. Pearson Education, Harlow.

Lieber, R. (2004). Morphology and Lexical Semantics. Cambridge: Cambridge University Press.

Plag, I. (2003). Word-Formation in English. Cambridge: Cambridge University Press.

Soanes, C. & Stevenson, A. (eds) (2004). Concise Oxford English Dictionary, 11th Edition. Oxford: Oxford University Press.

Internet sources
Bauer, L., 2008:
http://www.stuff.co.nz/blogs/opinion/735053/i-Language-is-embiggened-by-words-that-don-t-exist-i
British National Corpus. Ústav Českého národního korpusu FF UK, Praha. Available from WWW: <http://ucnk.ff.cuni.cz>.
Český národní korpus – SYN2000, 2000. Ústav Českého národního korpusu FF UK, Praha. Available from WWW: <http://ucnk.ff.cuni.cz>.
Facebook exchange: https://www.facebook.com/permalink.php?story_fbid=301321023242571&id=296963433678330
Leech et al., 2001: http://ucrel.lancs.ac.uk/bncfreq/lists/5_3_all_rank_adjective.txt
Urban Dictionary: http://www.urbandictionary.com/define.php?term=embiggen

• • •

Daniela Kolbe-Hanna
Trier University
kolbe@uni-trier.de

Variation in complementizer choice between *if* and *whether*

This paper examines the variation between *if* and *whether* as complementizers in interrogative subordinate clauses. The choice of a *wh*-word in open interrogative clauses is determined by semantic context (*I don't know when that happened*) and thus not variable, but in *yes-no* or closed interrogative clauses (*I'm not sure if / whether that happened*) speakers may choose between *if* and *whether*. In contrast to the abundance of research on the variation between retention and omission of the complementizer *that* (see, e.g. the overview in Szmrecsanyi & Kolbe-Hanna, to appear), there are, to my knowledge, only a few studies of the choice between *if* and *whether* in interrogative subordinateclauses. Zieglschmid (1929) and Steinbach (1929) describe prescriptive and futile efforts to promote the exclusive use of *if* as conditional subordinator and of *whether* for interrogatives. The overview in Biber et al. (1999: 690–693) shows that *if* is more frequent than *whether* in conversation and that *whether* is equally frequent across different registers. In a study of dialect data, Kolbe (2008, 129–136) finds that *if* is slightly more frequent overall and *whether* is more likely to be used in Northern British dialects, after the matrix verb *know* and by older speakers, and is less frequent in the speech of women.

This study analyses the variation between *if* and *whether* in data from ICE-GB, ICE-Ireland and ICE-New Zealand. It primarily seeks to determine the strongest predictors of the variation between *if* and *whether* by means of a mixed effects logistic regression analysis and a random forest model. Choosing a mixed effects logistic regression

means that speakers' idiolectal preferences will be factored into modelling the distribution, and the random forest model helps to pinpoint the factors most relevant to the distribution (see Tagliamonte and Baayen 2012 for an overview and comparison of different statistical tools). Drawing on the register categories of ICE (see http://ice-corpora.net/ice/design.htm) will allow for a more fine-grained register distinction than in Biber et al. (1999) which distinguishes relatively broadly between conversation, news, fiction and academic prose). This study's major aim is to determine the most crucial predictors of the variation between *if* and *whether* by including cognitive factors that have proven to influence the choice between the retention of the complementizer *that* and its omission. These are, amongst others, morphosyntactic features in matrix and embedded clauses, and length of the embedded clauses. It therefore also explores how important cognitive factors are when the choice is not between retention and omission of a complementizer, but between a more explicit (*whether*) and a less explicit (*if*) option (cf. Rohdenburg 1996 for the notion of explicitness).

By combining register, speakers' social background and cognitive factors as predictors of the choice between *if* and *whether*, this study will shed more light on the weighting of external and internal factors in making linguistic choices and the involvement of cognitive factors in complementizer choice in general.

References

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

Kolbe, D. (2008). Complement clauses in British Englishes (Unpublished doctoral dissertation).

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. Cognitive Linguistics 7 (2), 149-182.

Steinbach, R. (1929). On usage in English. American Speech 4, 161-177.

Szmrecsanyi, B. & Kolbe-Hanna, D. (forthcoming). New ways of analyzing dialect grammars: Complementizer omission in traditional British English dialects. In S. Grondelaers & R. van Hout (Eds.), New ways of analyzing syntactic variation. Berlin, Boston: de Gruyter Mouton.

Tagliamonte, S. & H. Baayen (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. Language Variation and Change, 24, 135–178.

Zieglschmid, A. J. F. (1929). "If" for "whether". American Speech 5 (1), 50-51.

• • •

Kim-Sue Kreischer
University of Nottingham
kimsue.kreischer@nottingham.ac.uk

Using collocations to analyse intertextual discourse relations

Interpretation of the world is closely connected to often opposing ideological viewpoints or Discourses on the same situation or event. Yet Discourses are complex and emerge out of intertextual or co-occurring relations (Sunderland, 2004). So far the study of Discourse relations has been neglected from a quantitative perspective, even though collocation analysis and recent work on collocation networks (Brezina, McEnery, & Wattam, 2015) may enable such an analysis.

Yet corpus methods alone are not ideal to determine the function of discourse relations and how they are connected by readers. Discourses still need to be understood, held, and negotiated cognitively. I argue that an approach attending to both textual and cognitive aspects (cf. Mahlberg et al., 2017; Stockwell & Mahlberg 2015) has great potential to contextualise our understanding of how Discourses are formed in relation to each other. To this end, I advocate for using research on attention, which connects corpus and cognitive linguistics. Attention is created through textual choices, by drawing attention to some Discourses over others, and actively searched for by the discourse participants to understand the (ideological) meaning of a text.

The corpus used in this paper is the Irish Abortion Debate Corpus (IADC), which comprises Irish online and offline newspaper articles from 2005 to 2016. I will focus on a sub-corpus of about 55,000 words from October 2012 to 2013. This year covers the death of Savita Halappanavar, who dies from complications of a septic miscarriage after her request for an abortion was denied. Halappanavar's death became a headline news story, leading to national and international protests, increased media discussion on abortion access, and ultimately had legal results with the creation of the *Protection of Life During Pregnancy Act 2013*. In this light, I will analyse the discursive and intertextual construction of the concepts 'woman' and 'mother' in relation to the term 'church' during this time period.

In order to locate specific co-occurring discourses, I wrote a Python script that counts which collocational Discourses (co-)occur in the sub-corpus. To do so, the programme takes a list of the manually categorised collocates of each term and identifies which categories are present in a span of 5 words for each of the analysed concepts. The notion of 'attention' is included into the qualitative analysis by using the figure/ground constellation and Cognitive Grammar (CG) (Langacker, 2008) to assess the relative contribution of a Discourse and its relation to other Discourses.

This paper adds to recent research programmes in corpus linguistics integrating cognitive and psycholinguistic research (e.g., Mahlberg, Conklin, & Bisson 2014) to answer questions about the link between text and cognition. Integrating cognitive analyses into corpus research is not just beneficial, but also necessary. By combining the two, we can address questions of Discourse conceptualisation and expression from the perspective of writer and reader and how persuasive their contribution may

be. A combined approach also problematises the issue of frequency and importance at the heart of corpus linguistics (cf. Hoey 2005).

References

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

Hoey, M. (2005). Lexical Priming: A New Theory of Words and Language. London: Routledge.

Langacker, R.W. (2008). Cognitive Grammar: A Basic Introduction. Oxford/New York: Oxford University Press.

Mahlberg, M., Conklin, K., & Bisson, M.-J. (2014). Reading Dickens's characters: textual patterns and their cognitive reality. Language and Linguistics, 23(4), 369-388.

Mahlberg, M., Stockwell, P., de Joode, J., Smith, C. & O'Donnell, M.B. (2017). CLiC Dickens - Novel Uses of Concordances. Corpora, 11(3), 433-463.

Stockwell, P., & Mahlberg, M. (2015). Mind-modelling with corpus stylistics in David Copperfield. Language and Literature, 24(2), 129-147.

Sunderland, J. (2004). Gendered Discourses. Basingstoke: Palsgrave Macmillan.

• • •

Lisa Lehnen
University of Würzburg
lisa.lehnen@uni-wuerzburg.de

Studying pragmatic variation in second-language varieties of English with the International Corpus of English – challenges and opportunities

Seminal models to categorise Englishes (Kachru 1988; Schneider 2007) have contributed immensely to the study of varieties of English, particularly to a systematic description of variety status and corresponding linguistic features. When looking at the research, however, it shows that the main, sometimes exclusive, focus lies on the morphosyntactic and lexical analysis of varieties. Seldom does the analysis integrate the pragmatic level, such as the study of speech act realisation strategies or interactional patterns.

There might be theoretical and methodological reasons for this: Within sociolinguistics, the study of variation in language structure has traditionally been emphasised, while pragmatic theories have largely sought to detect universal tendencies in language use (Schneider & Barron, 2008: 2-7). To merge these two paradigms, vari-

ational pragmatics provides a framework for the analysis of pragmatic variation in relation to micro- and macro-social factors within pluricentric languages (Schneider & Barron 2008: 15-22). Methodologically, it proclaims "the principles of empiricity, contrastivity and comparability" (Schneider 2010: 1030), which may, however, be hard to adhere to when studying second-language varieties of English due to the diverging impact of first languages and socio-cultural histories. The *ICE* project with "its principle aim […] to provide the resources for comparative studies of English" (Greenbaum 1996: 3) may, nonetheless, meet these requirements and thus provide the means to conduct analyses of pragmatic variation across Englishes.

One interesting interactional phenomenon to look at in this context is *disagreeing*, i.e. "the expression of a view that differs from that expressed by another speaker" (Sifianou 2012: 1554). It is highly context-sensitive and multifunctional in that it may be a sign of both conflict and solidarity between interactants. Needless to say, its function, frequency and the use of mitigation strategies (Brown & Levinson 1987; Spencer-Oatey 2008) also vary cross-culturally (see for instance Beebe & Takahashi 1989; Cheng & Tsui 2009). Therefore, due to its complex nature and cultural sensitivity, the particulars of *disagreements* are certainly worth investigating.

The present work-in-progress attempts to elucidate the challenges and opportunities that a large-scale corpus project, such as *ICE*, provides for the study of *disagreements* across second-language varieties of English. It sets out with a study of private conversations in *ICE* (among others ICE Hong Kong and ICE Jamaica) and seeks to identify similarities and differences in *disagreeing*. Furthermore, it discusses how representative of the respective culture the speech events in *ICE* are and if the data of different ICE corpora are comparable. Since metapragmatic information on speakers and setting is crucial for a pragmatic analysis, its availability for the *ICE* data will also be assessed. When the project was still in its infancy, Leitner observed that "the differences that accumulate on the pragmatic and discoursal levels […] do not seem to be catered for adequately" (1992, p. 33). Taking the richness of spoken data into account, though, *ICE* may still allow for theoretically and methodologically profound research of pragmatic variation.

References

Beebe, L. M. & Takahashi, T. (1989). Sociolinguistic variation in face-threatening speech acts: Chastisement and disagreement. In M. R. Eisenstein (Ed.), Topics in language and linguistics. The dynamic interlanguage. Empirical studies in second language variation. New York: Plenum Press, 199-218.

Brown, P. & Levinson, S. C. (1987). Politeness. Some universals in language use. Cambridge: Cambridge University Press.

Cheng, W. & Tsui, A. B. (2009). 'ahh ((laugh)) well there is no comparison between the two I think': How do Hong Kong Chinese and native speakers of English disagree with each other? Journal of Pragmatics, 41 (11), 2365-2380.

Greenbaum, S. (1996). Introducing ICE. In S. Greenbaum (Ed.), Comparing English worldwide: The International Corpus of English. Oxford: Clarendon Press, 3-12.

Kachru, B. B. (1988). The sacred cows of English. English Today, 4 (4), 3-8.

Leitner, G. (1992). International Corpus of English: Corpus design - problems and suggested solutions. In G. Leitner (Ed.), New directions in English language corpora. Methodology, results, software developments. Berlin, New York: Mouton de Gruyter, 33-64.
Schneider, E. W. (2007). Postcolonial English: Varieties around the world. Cambridge: Cambridge University Press.
Schneider, K. P. (2012). Appropriate behaviour across varieties of English. Journal of Pragmatics, 44 (9), 1022-1037.
Schneider, K. P., & Barron, A. (2008). Where pragmatics and dialectology meet: Introducing variational pragmatics. In K. P. Schneider & A. Barron (Eds.), Variational pragmatics. A focus on regional varieties in pluricentric languages. Amsterdam: John Benjamins, 1-32.
Sifianou, M. (2012). Disagreements, face and politeness. Journal of Pragmatics, 44 (12), 1554-1564.
Spencer-Oatey, H. (2008). Face, (im)politeness and rapport. In H. Spencer-Oatey (Ed.), Culturally speaking. Culture, communication and politeness theory (2nd ed.). London: Continuum, 11-47.

• • •

Xiaofei Lu
The Pennsylvania State University
xxl13@psu.edu

J. Elliott Casal
The Pennsylvania State University
jec368@psu.edu

Yingying Liu
The Pennsylvania State University
yzl222@psu.edu

A corpus-based study of the rhetorical functions of syntactically complex sentences in research article introductions

This paper investigates the rhetorical functions of syntactically complex sentences in research article (RA) introductions. The past decade witnessed a growing interest in the "integration of genre analysis and corpus-based investigations" (Flowerdew 2005: 5) in English for Academic Purposes (EAP) writing research. Much research in this area focused on identifying rhetorical moves that realize the communicative functions of different genres. Recent research has also started to attend to the links

between specific linguistic features and rhetorical moves. For example, some recent corpus-based studies have explored writers' use of formulaic language (Durrant & Mathews-Aydınlı 2011) and lexical bundles (Cortes 2013) in achieving the rhetorical moves strongly associated with academic research genres. Meanwhile, while many studies have quantitatively assessed the relationship of syntactic complexity to writing quality (e.g., Beers & Nagy 2009) and first and second language (L2) development (e.g., Lu, 2009, 2011), studies that adopt meaning-based approaches to examine the rhetorical functions of syntactically complex sentences in academic writing are scarce (e.g., Ryshina-Pankova 2015). In light of these contexts, the current study aims to move forward our understanding of how expert writers exploit syntactically complex sentences in accomplishing their rhetorical goals through meaning-based analysis of syntactic complexity.

Our data consists of the introduction sections of published RAs in the Corpus of Social Science Research Article Introductions (COSSRA) compiled by our research team. COSSRA includes 600 RAs published in 2012-2016 in six social science disciplines (Anthropology, Applied Linguistics, Economics, Political Science, Psychology, and Sociology), with 100 RAs sampled from five top journals in each discipline. The journals were selected according to impact factor and member checked with disciplinary experts to confirm representativeness. Four widely used operationalizations of syntactic complexity are explored here: mean length of sentence (MLS), left embeddedness (i.e., number of words before the main verb), amount of subordination (i.e., number of subordinate clauses per sentence), and nominalizations. The data will be analyzed using the L2 Syntactic Complexity Analyzer (Lu 2010), the D-Level Analyzer (Lu, 2009), and other tools. A sentence will be considered syntactically complex if it meets the threshold established for MLS, left embeddedness, amount of subordination, or nominalizations. The syntactically complex sentences will be manually categorized according to the Create A Research Space model (Swales 1990, 2004) of rhetorical moves for RA introductions. A quantitative analysis of the distribution of the four types of syntactically complex sentences across different rhetorical moves will be performed, complemented by a qualitative analysis of the rhetorical functions of sentences engaging multiple types of syntactic complexity. The implications of our findings for EAP writing research and pedagogy, genre analysis, and L2 writing syntactic complexity research will be discussed.

References

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? Reading and Writing: An Interdisciplinary Journal, 22 (2), 185-200.

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. Journal of English for Academic Purposes, 12, 33-43.

Durrant, P., & Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. English for Specific Purposes, 30, 58-72.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies. English for Specific purposes, 24, 321-332.
Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. International Journal of Corpus Linguistics, 14 (1), 3-28.
Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15 (4), 474-496.
Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. TESOL Quarterly, 45 (1), 36-62.
Ryshina-Pankova, M. (2015). A meaning-based approach to the study of complexity in L2 writing: The case of grammatical metaphor. Journal of Second Language Writing, 29, 51-63.
Swales, J. M. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge: Cambridge University Press.
Swales. J. M. (2004). Research Genres: Explorations and Applications. Cambridge: Cambridge University Press.

• • •

Claudia Lückert
University of Münster
c.lueckert@uni-muenster.de

Are some words more ‚proverbial' than others? The lexical profile of modern American English proverbs based on CAEP-data

When compared to other types of fixed expressions proverbs are characterised by a higher degree of linguistic complexity and a generally low token frequency (Moon 1998, Grzybek 2012). At the same time, however, infrequent proverbs can still be very familiar to speakers (Grzybek 2012: 107). This apparent discrepancy can be resolved by an assumed ‚back-up system' in the mental lexicon (in line with usage-based theory; see Bybee 2013) that strengthens the memory representation of proverbs (Lückert forthcoming). This hypothesis is tested in a study combining corpus-based analyses and psycholinguistic tasks and measurements. First, a corpus of modern American English proverbs (*CAEP - Corpus of American English Proverbs*) was compiled so as to extract the most frequent words used in proverbs and those that are characterised by a high contextual predictability. The corpus which includes more than 6,600 proverb variants draws on published material and includes further data from the *Corpus of Contemporary American English* (*COCA*) and the Web. *CAEP* integrates current corpus

data on the token frequency of proverbs (see Steyer 2015) and current experimental data on familiarity (Chlosta & Grzybek 2005, Haas 2008). Steyer calls for „the creation of new proverb collections or modern proverb information systems" which bring together various types of information (2015: 222). *CAEP* combines information on the form of proverb variants with data on frequency and familiarity. In this way the corpus may be of interest to a variety of researchers as it allows searching the linguistic structure of thousands of proverb variants (for example in view of linguistic structuring principles such as Panini's Principle or the distribution of conceptual metaphors). The corpus is compiled in a spreadsheet version with Microsoft Excel. Proverb items are only used for generating the lexical profile if a minimum frequency of occurence is reached in various general language corpora. The absolute frequency of all the individual words in the data set is identified and their contextual predictability is calculated with a log likelihood test. In addition to the log likelihood values an index is calculated on the basis of data from both *CAEP* and *COCA*. The paper reports on the lexical profile of modern American English proverbs and argues that some words are indeed ‚more proverbial' than others - that is quite a number of words are strongly associated with the proverb as a category.

References

Bybee, J. (2013). Usage-based theory and exemplar representation of constructions. In T. Hoffmann & G. Trousdale (Eds.), The Oxford Handbook of Construction Grammar. Oxford: Oxford University Press, 49-69.

Chlosta, Ch. & Grzybek, P. (2005). Varianten und Variationen anglo-amerikanischer Sprichwörter – Dokumentation einer empirischen Untersuchung. ELiSe: Essener Linguistische Skripte elektronisch, 5 (2), 63-145.

Grzybek, P. (2012). Facetten des parömiologischen Rubik-Würfels. Kenntnis = Bekanntheit? In K. Steyer (Ed.), Sprichwörter multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömiologie. Tübingen: Gunter Narr, 99-138.

Haas, H. A. (2008). Proverb familiarity in the United States: Cross-regional comparisons of the paremiological minimum. Journal of American Folklore, 121, 319-347.

Lückert, C. (forthcoming). A psycholinguistic approach to the conventionalisation and variation of proverb structure. In N. Filatkina & S. Stumpf (Eds.), Conventionalization and variation. Konventionalisierung und Variation. Frankfurt a.M.: Peter Lang.

Moon, R. (1998). Fixed expressions and idioms in English. A corpus-based approach. Oxford: Clarendon.

Steyer, K. (2015). Proverbs from a corpus-linguistic point of view. In H. Hrisztova-Gotthardt & M. A. Varga (Eds.), Introduction to Paremiology. A Comprehensive Guide to Proverb Studies. De Gruyter Open, 206-228.

• • •

Krittaya Ngampradit
Kasetsart University
krittaya.ngampradit@gmail.com

Raksangob Wijitsopon
Chulalongkorn University
rwijitsopon@gmail.com

A Corpus-based Study of Metadiscoursal Boosters in Applied linguistics Dissertations in Thailand and in the United States

It has been suggested that academic writing is often influenced by such socio-cultural context as academic settings and the writer's perception of identities and stances. These discourse-pragmatic factors can be manifested through linguistic choices and patterns found in the text. Based on this assumption, this pilot study investigates the use of „boosters", markers of certainty and authorial commitment to propositions, in two corpora of applied linguistics doctoral dissertations:

(1) a corpus of 20 PhD dissertations submitted to universities in Thailand
(2) a corpus of 20 PhD dissertations submitted to universities in the USA

While previous studies on boosters, metadiscoursal devices and academic discourse tend to focus on disciplinary variation (e.g. Bondi 2008), the present study, as reflected by the two corpora involved, addresses cross-cultural and rhetorical-chapter variations. Using Hyland's (2005) interactional model of metadiscourse, different boosters in the dissertations were identified and classified. The analysis of 40 randomly selected writings from universities in two different countries reveals that there are significant discrepancies in terms of distribution and usage patterns of metadiscoursal boosters across the corpora e.g. dissertations by Thai student writers displayed more substantial use of metadiscoursal boosters while a more limited range of structural patterns where boosters were utilized was evident in their writings; boosters were found to occur most frequently in Results, Literature Review and Discussion chapters, respectively; and the most frequently used categories of metadiscoursal boosters found in the three chapters were verbs, adverbs and adjectives, correspondingly. The differences are argued to reflect constructions of student writers' identities and stances, which are in turn linked to specific cultural and institutional settings in which the writing is produced as well as their readership. This qualitative interpretation of the rhetorical differences is made in the light of genres, part-genres, discourse communities and writing practice. The study concludes with some pedagogical implications for academic writing in the EFL context.

References

Bondi, M. (2008). Emphatics in Academic Discourse: Integrating Corpus and Discourse Tools in the Study of Cross-disciplinary Variation. In Corpora and Discourse: The Challenges of Different Settings (pp.31-55). Amsterdam: John Benjamins.

Hyland, K. (2005). Metadiscourse: Exploring Interaction in Writing. London: Bloomsbury.

• • •

Yoshiyuki Notohara
Doshisha University
ynotohar@mail.doshisha.ac.jp

Exploring the epistemic associations between tense-aspect form-meaning usage patterns and canonical event schemata in the spoken English corpus: A collostructional analysis

This study explores the epistemic associations between five tense-aspect (TA) form-meaning usage patterns (e.g., *-s/-es* + immediate reality-present tense) and thirteen canonical event schemata (e.g., States) through the tagged BNC spoken component (2007) from Sketch Engine. It also clarifies the canonical TA form-meaning usage patterns of canonical event schemata through each most frequent verb (e.g., *be*). Finally, it discusses and specifies potential pedagogical applications of canonical event schemata in L2 grammar instruction as "minimal essentials."

In cognitive grammar, Langacker (1991; 2008) explains epistemic meaning aspects of TA usage patterns in English into four realities from the perspective of **clausal grounding** (establishing a basic connection between the interlocutors and the content evoked by a nominal or a finite clause): known reality, immediate reality, projected reality, and potential reality. Furthermore, Radden & Dirven (2007: 173) exemplifies the relationships between four realities and TA usage patterns as follows: (i) known reality-past tense (e.g., *Bill and Jane got married last year.*); (ii) immediate reality-present tense (e.g., *Bill and Jane are getting married today.*); (iii) projected reality-future tense (e.g., *Bill and Jane will get married next week.*); and (iv) potential reality-modal verbs (e.g., *Bill and Jane may be getting married soon.*). Thus, referring to Langacker's epistemic reality conception (2011: 67-73), new approaches to TA form-meaning usage patterns in L2 grammar instruction could be explored while figuring out embodied natures of TA form-meaning usage patterns in English.

A previous pilot study (Notohara 2015) conducted a collostructional analysis (e.g., Stefanovitsch & Gries 2003; Gries, & Stefanovitsch 2004; Gries 2011; Schmidt & Küchenhoff 2013) and confirmed the association strength between five TA form-meaning usage patterns and thirteen canonical event schemata through the grammatically-tagged ICE-GB R2 (2006) spoken component. As a result, the following three associations emerged:

(1) occurrence schema (e.g., States) is mainly related to immediate reality-present tense.
(2) process schema (e.g., Processes) and psychological schema (e.g., Emotion) are mainly related to known reality-past tense.
(3) force-dynamic schemata (e.g., Action) are mainly related to both projected reality-future tense and potential reality-modal verbs.

However, it is still difficult to recognize the three epistemic associations as 'canonical' because the ICE-GB R2 corpus is small and its register is relatively academic. Besides, tense form-meaning associations are clarified to some extent; in contrast, aspect ones are still unclear. Therefore, further related investigations are needed through a well-balanced larger spoken corpus. This study reconfirms the three canonical associations as follows: (i) raw frequencies of TA verb inflectional morphemes (e.g., *-s/-es*) of thirteen canonical verbs (e.g., *be*) in the spoken component of the tagged BNC are respectively described referring to CLAWS5 tags; (ii) thirteen canonical verbs are grouped through the correspondence analysis to confirm their distributional similarities; (iii) the association strength between TA usage patterns and thirteen canonical verbs are confirmed through collostructional analysis.

Finally, potential pedagogical applications of the newly observed canonical associations to EFL learners in usage-based L2 grammar instruction are discussed in the following area: construction frequency effects on learners' interlanguage development, pedagogical task design, supplementary explicit construction instruction.

References

Bergen, B. & Wheeler, K. (2010). Grammatical aspect and mental simulation. Brain & Language, 112, 150-158.

Ellis, N. (2013). Frequency-based grammar and the acquisition of tense and aspect in L2 learning. In R. Salaberry, & L. Comajoan (Eds.), Research Design and Methodology in Studies on L2 Tense and Aspect. (pp.89-117). Berlin: Walter de Gruyter.

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

Gries, S.Th. (2011). Corpus data in usage-based linguistics. In M. Brdar, S. Th. Gries, & M, Fucks. (Eds.), Cognitive Linguistics: Convergence and Expansion (pp.237-256). Amsterdam: John Benjamins.

Langacker, R.W.(1991). Foundations of Cognitive Grammar, Vol. 2: Descriptive Application. Stanford, CA: Stanford University Press.

Langacker, R.W. (2008). Cognitive Grammar. Oxford: Oxford University Press.

Langacker, R.W. (2011). The English present: Temporal coincidence vs. epistemic immediacy. In A. Patard & F. Brisard (Eds.), Cognitive Approaches to Tense, Aspect, and Epistemic Modality (pp. 45-86). Amsterdam: John Benjamins.

Niemeier, S. (2013). A cognitive grammar perspective on tense and aspect. In R. Salaberry, & L. Comajoan. (Eds.), Research Design and Methodology in Studies on L2 Tense and Aspect. (pp.11-56). Berlin: Walter de Gruyter.
Notohara, Y. (2015). Tense and aspect usage patterns of canonical verbs in the spoken English Corpus. Paper presented at the British Association of Applied Linguistics 48th Annual Conference, Birmingham, UK.
Radden, G. & Dirven, R. (2007). Cognitive English Grammar. Amsterdam: John Benjamins.
Schmidt, H. & Küchenhoff, H. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinings. Cognitive Linguistics, 24 (3), 531-577.
Stefanowitsch, A. & Gries, T. (2003). Collostructions: Investigating the interaction of words and constructions. International Journal of Corpus Linguistics, 8 (2), 209-243.

• • •

Malgorzata Paprota
Maria Curie-Sklodowska University, Lublin, Poland
m.paprota@gmail.com

(Not) helping. Role allocation in the representation of the British welfare state across the British press: a corpus-assisted discourse analysis

The British welfare state can hardly be considered a stable concept. Some definitions take a broader view of the concept as "a mix of universal and comprehensive policies through which government became more positively responsible for the promotion of individual welfare" (Jones and Lowe 2002, 6). More narrowly, it is understood as a system of social security benefits. More radically, to quote a historian of the welfare state, "[a]s an entity it does not exist – it is a collection of services and policies and ideas and taxes, including tax reliefs, whose boundaries expand and contract over time" (Timmins 2001, 7). Motivated by this instability, this proposed paper seeks to outline the roles allocated to the welfare state as a social actor, and the conceptual model of the welfare state they entail, across a corpus of British press texts.

The corpus, about 830,00 tokens in size, consists of four newspaper subcorpora: two British conservative newspapers, the Daily Mail and the Daily Telegraph, and two left-leaning ones, the Guardian/Observer and the Daily Mirror. The corpus has been obtained from the Lexis Nexis database and comprises texts where the search term 'welfare state' occurs twice to maximise topicality. The timeframe (January 2008-April 2015) has been chosen to coincide with the economic crisis, which increasingly af-

fected government policy with regard to the welfare state, and includes two election campaigns.

Positioned within corpus-assisted discourse studies (see Partington, Duguid, and Taylor 2013), this proposed paper analyses the verbal collocates of the term 'welfare state' obtained from Sketch Engine with recourse to the allocation of semantic roles, an element of van Leeuwen's network of the representation of social actors (2008), in order to outline the patterns of agency. (The assumption is that an institution or an abstract entity may be constructed as having agency and so social actors need not be immediately human.) The paper finds that the range of roles attributed to the welfare state is fairly limited across the corpus, where the bulk of roles are generalisable as Actor in the material action processes of helping or damaging, with the conceptual models of the welfare state generalisable as the safety net, the piggy bank, the springboard, and the trap. The political affiliations of the newspapers are reflected in the modalities ascribed to the roles rather than in the roles themselves: the 'helper' role tends to be counterfactual (or have a past reference) in the conservative side of the corpus, while the 'destroyer' role is represented as factual; the reverse tends to be the case for the left-leaning side of the corpus, although the presence of the counterfactual 'helper' is substantial.

References

Jones, M. & Lowe, R. (2002). From Beveridge to Blair. The First Fifty Years of Britain's Welfare State 1948-98. Manchester and New York: Manchester University Press.

Partington, A., Duguid, A. & Taylor, C. (2013). Patterns and Meanings in Discourse. Theory and practice in corpus-assisted discourse studies. Amsterdam: John Benjamins.

Timmins, N. (2001). The Five Giants. A Biography of the Welfare State. 2nd ed. London: HarperCollins

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

• • •

Pam Peters
Macquarie University, NSW Australia
pam.peters@mq.edu.au

Tobias Bernaisch
Justus-Liebig University, Giessen Germany
Tobias.J.Bernaisch@anglistik.uni-giessen.de

Kathleen Ahrens
Hong Kong Polytechnic University
kathleenahrens@gmail.com

Modality, rhetoric and regionality in corpora from Greater China

This paper is associated with the *VARIETIES OF ENGLISH IN THE INDO-PACIFIC (VEIP)* project, focusing on socio-cultural and linguistic change in contemporary Hong Kong, China and Taiwan. It also reflects the conference theme *Corpus et Orbis,* using matched regional corpora to examine the world views reflected in English-language news sources from three "expanding circle" countries.

This research builds on previous research on political metaphors that reflect socio-cultural change in Hong Kong, as used in the rhetoric of its leaders before and after the administrative changeover in 1997 (Ahrens 2016). Other linguistic vehicles highly sensitive to the socio-political positioning of the speaker or writer are (i) modal verbs and (ii) the uses of personal pronouns. In research on the language of British election manifestos by the major political parties, Rayson (2008) demonstrated the "keyness" of *would* by the opposition party (Liberal Democrats), and of *our* by the governing Labour Party. Regular use of the first and second person pronouns, contribute to the interactive dimension of communication (Biber 1988); and increasing use of the quasi-modals instead of the core modals (Leech & Mair 2009, Collins 2009) suggests the desire to give a more colloquial tenor to the discourse, and reduce the communicative distance between speaker/writer and the audience. While recent research on modal usage in Hong Kong (Noel & Van der Auwera 2014; Loureiro-Porto 2016) provide useful benchmarks on the relative frequencies of the two types, there has been no comparable research on their frequencies in data from Mainland China and Taiwan. The further question of how the use of modals contributes to the positioning of English-medium political rhetoric in these different regions of Greater China also invites research.

The Greater China English Corpus used in this study was compiled by Ahrens from editorials/commentaries in English-language newspapers in the three regions. It includes copy from two newspapers in the People's Republic of China (PRC; *China Daily* and *Global Times*) during the period from 2011–2016 (61,904 words); from three news sources in Hong Kong (*South China Morning Post*, *The Standard* and the online

Hong Kong Free Press) from 2012–2016 (51,216 words); and from two newspapers in Taiwan (*The China Post* and *Taipei Times*) from 1999–2016 (33,761 words).

The frequency patterns of the modals/quasi-modals proved to be significantly different in each of the three regional subcorpora. In the PRC, the most frequent modals by far were *will* (expressing strong prediction) and *should* (moderately strong obligation), along with comparatively low use of the quasi-modals. Combined with low usage of interactive pronoun, the PRC texts project the voice of authority offering little consultative space, in line with its dominant political position in Greater China. Whereas in both Hong Kong and Taiwan, the frequency of *will* was closely matched by *would* (the more hypothetical modal), with quasi-modal *going to* providing a more informal expression of future expectations. Combined with more frequent use of the interactive pronouns, the tenor of the Hong Kong and Taiwanese texts is less assertive and more collaborative, in keeping with their more fluid status within Greater China for different historical reasons. Further multifactorial analysis is being carried out on the distributions of the modals/quasi-modals to embrace other linguistic factors such as analytic and contracted negation; animacy of the agent (cf. Deshors & Gries 2014); degree of formality (cf. Nini 2014); and regional provenance. The data will be modelled complementarily via conditional inference trees (Hothorn et al. 2006), and linear regression models in R (R Core Team 2016). It will provide further evidence of how socio-political developments are reflected in linguistic differences in the lexicogrammars of these three Expanding-Circle countries.

References

Ahrens, K. (2016). Contrasting BUILDING metaphors in the Corpus of Political Speeches. Paper presented at International Computer Archive of Modern & Medieval English (ICAME) 37, Chinese University of Hong Kong, Hong Kong.

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

Collins, P. (2009). Modals and quasi-modals in World Englishes. World Englishes 28(3), 281–292.

Deshors, S. C. & Gries, S. Th. (2014). A case for the multifactorial assessment of learner language: The uses of may and can in French-English interlanguage. In D. Glynn & J. A. Robinson (Eds.), Corpus methods for semantics: Quantitative studies in polysemy and synonymy. Amsterdam: John Benjamins, 179–204

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15(3), 651–674.

Loureiro-Porto, L. (2016). (Semi-)modals of necessity in Hong Kong and Indian Englishes. In E. Seoane & C. Suárez-Gómez (Eds.), World Englishes: New theoretical and methodological considerations. Amsterdam: John Benjamins, 143-172.

Nini, A. (2014). Multidimensional analysis tagger 1.2 – Manual. Retrieved from: http://sites.google.com/site/multidimensionaltagger.

Noël, D. & van der Auwera, J. (2014). Recent quantitative changes in the use of modals and quasi-modals in the Hong Kong, British and American printed press.

In P. Collins (Ed.), Grammatical Change in English World-Wide (pp. 437–464). Amsterdam: John Benjamins.
R Core Team. (2016). R: A Language and environment for statistical computing. Retrieved from R Foundation for Statistical Computing, Vienna, Austria: https://www.R-project.org/
Rayson, P. (2008). From keywords to key semantic domains. International Journal of Corpus Linguistics 13(4), 519–549.

• • •

Nele Põldvere
Lund University
nele.poldvere@englund.lu.se

Carita Paradis
Lund University
carita.paradis@englund.lu.se

Victoria Johansson
Lund University
victoria.johansson@ling.lu.se

The London-Lund Corpus 2: A new corpus of spoken British English in the making

This methodological paper describes and critically examines the major stages of building a spoken language corpus by reporting on the process of compiling the London-Lund Corpus 2 (LLC 2) of spoken British English. LLC 2 is of the same size and compiled on the same principles as its predecessor, the London-Lund Corpus (LLC 1), launched in 1975 (Svartvik 1990). LLC 2 will allow linguists to study contemporary speech in different contexts and in combination with LLC 1 it will also allow us to study language change in a principled way using a comparable data set of the language spoken 50 years ago.

While LLC 2 shares important traits with LLC 1, efforts are made to also make it synchronically representative (see Leech, 2007 for the incompatible relation between comparability and representativeness). On the one hand, LLC 2 is similar to LLC 1 in that priority is given to spontaneous face-to-face conversation, the most basic type of language use (Clark, 1996), and data retrieved from public resources (e.g. broadcast interviews and parliamentary debates). On the other hand, LLC 2 differs from LLC 1 in that it includes computer-mediated communication, more specifically, Skype con-

versations, in order to represent speech situations that use technologies characteristic of the 21st century.

The compilation of LLC 2 entails the completion of four fundamental stages (modified from Thompson 2004):

1. Data collection (recordings of spoken communication)
2. Transcription of the recordings
3. Markup and annotation
4. Access to the corpus

First, similar to LLC 1, the majority of the spontaneous face-to-face conversations are recorded at the University College London with native speakers of British English. Detailed information about the speakers is obtained in order to support sociolinguistic analyses of the data. At the transcribing stage, the recordings are turned into written form following a detailed transcription scheme. The scheme is largely based on the International Corpus of English; however, a number of modifications have been made. For instance, transcriptions in LLC 2 also include timestamps that connect each speaker turn to the corresponding location in the audio file, allowing for prosodic analyses of the conversations. Stage 3 entails the computerization of the transcriptions. Hardie's (2014) *Modest XML for Corpora* is followed in the encoding of the corpus for the purposes of distribution and archiving. Furthermore, in contrast to LLC 1, anonymisation is carried out not only in the transcriptions but also in the audio files themselves by altering the tonal patterns of names. This means that the audio files will be made available to the public alongside the timestamped transcriptions from the Lund University Humanities Lab's server (Stage 4).

Research on language variation and change is crucial for our understanding of the forces, motivations and mechanisms that languages are constantly subject to in communication. Without having access to large and modern computerized language resources, and especially to spoken language data in which change features prominently, these endeavours cannot be pursued. The compilation of LLC 2 will fill this gap and facilitate principled research in the field.

References

Clark, H. (1996). Using Language. Cambridge: Cambridge University Press.

Hardie, A. (2014). Modest XML for corpora: Not a standard, but a suggestion. ICAME Journal, 38, 73-103.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), Corpus Linguistics and the Web. Amsterdam: Rodopi, 133-149.

Svartvik, J. (Ed). (1990). The London-Lund Corpus of Spoken English: Description and Research. Lund: Lund University Press.

Thompson, P. (2004). Spoken language corpora. In M. Wynne (Ed.), Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books.

• • •

Ute Römer
Georgia State University
uroemer@gsu.edu

The role of formulaic sequences in the L2 acquisition of high-frequency English verbs

This work-in-progress report presents findings from a subcomponent of a new large-scale corpus-based study on the development of verb patterns in second language learners of English. Inspired by pioneering existing work on emerging constructions in first language (L1) acquisition (especially Tomasello 2003), and using methods from Corpus Linguistics and Natural Language Processing, the larger study aims to investigate how verb-argument constructions (e.g. the 'V n n' or ditransitive construction) emerge in the writing of second language (L2) learners of English at different levels of proficiency. English L1 acquisition studies have shown how children acquire a language by picking up on recurring patterns in their input, hence building a mental constructicon (e.g., Goldberg, Casenhiser & Sethuraman 2004; Tomasello, 1992). Also, Tomasello (2003) demonstrated how in L1 acquisition children move from simple formulaic constructions to more complex ones.

Focusing on L2 instead of L1 acquisition, the current study discusses whether, and if so in what ways learners develop abstract, variable constructions from fixed, memorized chunks. We will address the following two research questions: (1) What role do formulaic sequences play in the L2 acquisition of verb constructions? and (2) Are there significant observable differences in the L2 acquisition of verb constructions between L1 German and L1 Spanish learners? To address these questions, we exhaustively extract data on verbs and the constructions they occur in from a cross-sectional corpus of L2 learner writing. The corpus is a 6-million word subset of EFCAMDAT, the Education First-Cambridge Open Language Database (Geertzen, Alexopoulou, & Korhonen, 2013), consisting of over 68,000 texts produced by German (dominant L1 German) and Mexican (dominant L1 Spanish) learners at CEFR levels A1 through C1. Using the software *Collocate* (Barlow, 2015), we are currently extracting recurring multi-word clusters (spans 3, 4, and 5) around the 50 most frequent verbs in EFCAMDAT, together with information on frequency and cluster association strength (Mutual Information). For each L1-proficiency level combination, this will give us verb sequences that are particularly frequent and/or particularly fixed.

In order to determine what role formulaic language plays in L2 acquisition, we are especially interested in tracing the development of fixed, formulaic verb sequences from low to higher proficiency levels. In our presentation, we will share selected results on verb construction development across learner proficiency levels and comment on similarities and differences between L1 German and L1 Spanish learners. We expect to find predominantly fixed sequences at beginning levels which then become more flexible and productive as learners progress to more advanced levels. The resulting findings are likely to expand our understanding of the processes that underlie construction acquisition in an L2 context.

References

Barlow, M. (2015). Collocate (Version 2.0) [Computer Software]. Houston, TX: Athelstan.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). Proceedings of the 31st Second Language Research Forum (SLRF). Carnegie Mellon University: Cascadilla Press.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. Cognitive Linguistics, 15, 289-316.

Tomasello, M. (1992). First Verbs. A Case Study of Early Grammatical Development of Cognition and Action. Cambridge, MA: MIT Press.

Tomasello, M. (2003). Constructing a Language. A Usage-based Theory of Language Acquisition. Cambridge, MA: Harvard University Press.

• • •

Julia Schlüter
University of Bamberg
julia.schlueter@uni-bamberg.de

Günter Rohdenburg
University of Paderborn
rohdenburg@onlinehome.de

Prosodic salience as a determinant of morphological marking

The purpose of this paper is to provide evidence for a widespread cross-linguistic correlation between the relative prosodic salience of a given word and its degree of morphological markedness. Exploring relevant variation phenomena from both Middle and Present-Day English, the paper illustrates the effects of three kinds of prosodic prominence: sentence-stress, word-stress and the special case of stranded prepositions, which necessarily feature strong forms.

The first two examples come from Middle English. One is the loss of the old dative inflection in the infinitive (e.g. *to donne*), which occurred in early Middle English, compare (1a). The other is the spread of the *-s*-suffix that comes to mark possessive pro-forms of the type *hires*, *oures*, *youres* and *theires* around the thirteenth century, as illustrated in (2a).

(1) a. Alle ðo þing ðe ðu hauest te **dónne**, do it mit ræde; (HC *Vices and Virtues*)
b. for ofte we weneð for to **don** alutel uuel …. (PPCME2, *Ancrene Riwle*)

(2) a. …so will I do ffor Godes sayke and **hírs**. (HC, *Dame Sirith*)
b. His herte **hire** wes alon, …. (HC, E. Stonor, *Letters*)

Analyses drawing on a large corpus of Middle English (Helsinki Corpus, PPCME2) have shown that a word occurring under sentence stress accommodates an optional additional morpheme better than a word in sentence-unstressed position. As a result, inflected infinitives are preserved better in examples like (1a) and given up earlier in examples like (1b). Conversely, suffixless possessive pro-forms tend to persist longer in examples like (2b) than in examples like (2a).

The variation phenomena in Present-Day English involve the two plural forms of the root *scarf* in (3) and the alternation of morphologically simple prepositions and their strengthened counterparts as in (4). The variant *scarves*, which carries double morphological plural marking, has – over the last few centuries – all but supplanted the less marked variant *scarfs*. By contrast, the morphologically complex preposition *upon* has increasingly lost ground to its simple rival *on*.

(3) a. …a collection of héadscar**fs** to rival the Queen's … (*Sunday Times* 1991)
b. …the colourful scár**ves** on display. (ibid.)
(4) a. I pounced **on** all things British, … (*The Times* 2004)
b. …every new theory that Mr Jencks pounces **upón**. (*The Times* 1997)

In line with the correlations seen in (1) and (2) – and on the basis of a large collection of British and American newspapers – the following tendencies have emerged: In (3), the prosodically less prominent final constituent in the compound *headscarfs* is likely to preserve the (morphologically) regular and shorter variant better than the corresponding simplex, and in cases like (4), it is the "qualitative prominence" (Cruttenden 2008: 268) associated with stranded prepositions that may be an important factor favouring the use of strengthened alternatives such as *upon* and *into*.

In addition, we will adduce some close parallels from English (Ciszek 2002: 122-127) and other West-Germanic languages in support of the generality of the correlation and discuss its functional motivation and delimitation from end-weight phenomena (Eitelmann 2016, Mondorf 2009: 99-107).

References

Ciszek, E. (2002). ME -lich(e)/-ly. Studia Anglica Posnaniensia, 38, 105-129.

Cruttenden, A. (Ed., 2008). Gimson's Pronunciation of English. 7th ed. London: Hodder.

Eitelmann, M. (2016). Support for end-weight as a determinant of linguistic variation and change. English Language and Linguistics, 20, 395-420.

Mondorf, B. (2009). More Support for More-Support: The Role of Processing Constraints on the Choice between Synthetic and Analytic Comparative Forms. Amsterdam/New York: Benjamins.

• • •

Irma Taavitsainen
University of Helsinki
irma.taavitsainen@helsinki.fi

Gerold Schneider
University of Konstanz/ University of Zurich
gschneid@ifi.uzh.ch

Topics of medical writing in Early Modern and Late Modern English medical texts 1500-1800

Our presentation will report on an empirical corpus-driven study on changes of topics in medical writing in three hundred years. This is a pilot study applying a method of digital humanities to tracing the foci of medical writing from scholasticism to empiricism and to "enquiry as a thought style". The outlines are presented in histories of medicine, but linguistic studies give us more accurate knowledge about changes in different layers of writing. Our aim is to achieve an overall view of the various developments, to interpret equally the linguistic, historical, and social world through corpora.

The data comes from two corpora: Late Modern English Medical Texts 1700-1800 (LMEMT, fc) and Early Modern English Medical Texts 1500-1700 (EMEMT, 2010); together they contain over four million words and yield a diachronic perspective from the beginning of the printed history of medical texts to the dawn of more modern approaches to medicine. They contain systematically collected materials with a full scale of texts and genres from academic writing to texts targeted at lay readerships as well as selections from both general and specialized periodicals. The two corpora represent different periods and cultures in the history of English, and our study takes context in all its aspects from the narrow linguistic context to discourse and genre context to the broad cultural context into account, and we also use a context-based, distributional approach.

Distributional methods go back to the ideas of Firth and de Saussure, and have been used in collocation research, lexical semantics and content analysis. We use a data-driven, bottom-up corpus-linguistic method to detect clusters of words that tend to co-occur in the same documents, the distributional method of Topic Modeling (Blei 2012), which allows us to step from words to concepts. As tool, we use MALLET (http://mallet.cs.umass.edu/topics.php) and collocation detection The text passages and concepts relevant to our research aims are located by the program, and allow us to detect new concepts, shifts in concepts, and linguistic and stylistic change. These are analyzed qualitatively, taking the sociolinguistic parameters of authors' education and levels of audience into account. Special attention will be paid to meaning-making practices, especially as previous research has shown that features of scholastic writing acquire new context-dependent meanings in lay texts for heterogeneous audiences after the heyday of the thought style.

References

Blei, D. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Early Modern English Medical Texts 1500-1700 (EMEMT, 2010). Compiled by I. Taavitsainen, P. Pahta, T. Hiltunen, M. Mäkinen, V. Marttila, M. Ratia, C. Suhr & J. Tyrkkö. Amsterdam: John Benjamins. CD_ROM published with a book.

Late Modern English Medical Text 1700-1800 (LMEMT, forthcoming). Compiled by I. Taavitsainen, T. Hiltunen, V. Marttila, P. Pahta, M. Ratia, C. Suhr & J. Tyrkkö. Amsterdam: John Benjamins.

• • •

Cathleen Waters
University of Leicester
cw301@le.ac.uk

Nick Smith
University of Leicester
ns359@le.ac.uk

***Very popular at that time*: Degree adverbs in the British radio chat show**

This paper reports on a study of real-time variation and change among degree adverbs in a specific register, namely the radio chat show, using a corpus of recordings of the BBC Radio 4 programme *Desert Island Discs* that stretches from the 1960s to the early 2000s. Degree adverbs can occur in a variety of contexts but often modify adjectives as illustrated in this example from our data (see further, below):

(1) „the first act was very brilliant . it was really Strindbergian . it was so good"

A search of the BNC indicates that broadcast interviews, to which chat shows belong, are one of the most prolific registers for degree adverb use. Degree adverbs have been examined in British standard and regional varieties using the methods of corpus linguistics (e.g. including a consideration of register, cf. Anderson 2006; Xiao and Tao 2007) and quantitative sociolinguistics (e.g. considering demographic characteristics of the speakers, cf. Barnfield and Buchstaller 2010; Ito and Tagliamonte 2003; Macaulay 2006). This paper brings together the two strands by considering social characteristics of the speakers in a radio chat show in British English.

Methodologically, our paper also builds on an earlier study (Smith and Waters, in prep.) of the *Desert Island Discs* corpus in which findings on grammatical variation and change were assessed and compared according to two sampling methods: i) ran-

dom sampling and ii) sociolinguistic judgment sampling (with approximately 105,000 words in each). That study identified patterns of variation and change in grammatical usage by an inductive approach, namely, identifying through Wmatrix (Rayson 2008) Key part-of-speech-tags in the respective subcorpora. The two approaches identified similar changes at register-level, but the sociolinguistic subcorpus allowed a better insight into how the changing demographics of the guests interacted with changes in the register. The present study uses the same dataset, but, by contrast, takes a micro-variation approach, by selecting a functional domain (degree adverbs) and analysing the linguistic and social factor groups that correlate with choices for that function.

Provisional findings are based on 1062 adjectives in the 1960s and 1188 instances in the 2000s, from the sociolinguistic subcorpus. The results indicate that the nature and direction of real time change in the corpus is in keeping with the apparent-time sociolinguistic findings, namely that *very* is the most common degree modifier across all time periods, but is losing ground to *really*. However, the rate at which this change is taking place is slower in our corpus than what has been observed in studies based on other registers. One possibility is that conventions for the public performance of talk (Bell and van Leeuwen 1994) in *Desert Island Discs* have been relatively resistant to change. Alternatively, the findings may be a result of demographics of the guests. Both of these explanations will be examined in our talk.

References

Anderson, W. (2006). „Absolutely, totally, filled to the brim with the Famous Grouse": Intensifying adverbs in SCOTS. English Today, 87: 22 (3), 10-16.

Barnfield, K. & Buchstaller, I. (2010). Intensifiers on Tyneside: longitudinal developments and new trends. English World-Wide, 31, 252-87.

Bell, P. & van Leeuwen, T. (1994). The Media Interview: Confession, Contest, Conversation. Kensington: University of New South Wales Press.

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

Macaulay, R. (2006). Pure grammaticalization: The development of a teenage intensifier. Language Variation and Change, 18, 267-83.

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

Xiao, R. & Tao, H. (2007). A corpus-based sociolinguistic study of amplifiers in British English. Sociolinguistic Studies 1 (2), 241-73.

• • •

Nuria Yáñez-Bouza
University of Vigo
nuria.y.b@uvigo.es

Victorina González-Díaz
University of Liverpool
V.Gonzalez-Diaz@liverpool.ac.uk

he neaded the money to finenance his biness – A new corpus of schoolchildren's writings before the UK National Curriculum

This paper reports on a newly developed corpus for the study of schoolchildren's writing in the recent educational history of the UK: *The APU Writing and Reading Corpus 1979-1988*.

The explicit teaching of English grammar in UK schools has been a topic of debate since the early twentieth century (Hudson & Walmsley 2005). The 1970s, in particular, witnessed the rise of concerns about 'literacy standards', which crystallised, for instance, in the Department of Education and Science setting up the *Assessment of Performance Unit* (APU, 1975) for monitoring national standards of performance through a series of language surveys. Based on these language surveys (Foxman et al. 1991), we have developed the APU Corpus within the framework of Educational Linguistics in order to explore pupils' writings before the establishment of the National Curriculum (1989) in the UK education system.

Recent research has demonstrated the potential of corpus linguistics as a solid aid in children's understanding of how language works (Sealey & Thompson 2004, 2006). However, the availability of data from the UK is still somewhat limited. Most corpora are either based on a small number of schools, are synchronic in nature, or focus on the post-National Curriculum era (cf. *Lancaster Corpus of Children's Writing*, *The Oxford Children's Corpus*, the *'Grammar for Writing'* project at Exeter); on the other hand, longitudinal corpora are, unfortunately, not publicly available in electronic format (cf. *Aspects of Writing in 16+ English Examinations between 1980 and 2014* by Cambridge Assessment). The APU Corpus contributes to this growing body of data with the added value that it provides a large set of historical materials which are linguistically annotated and which can be accessed via an online interface.

In this paper we describe the contents and methodology procedure of the corpus:

(I) *Data selection*. The corpus is divided into two major components: writings *by* children ('school scripts') and writings *for* children ('basal readers'). The school scripts are written by 11-year-olds taking Year-6 level of primary schools, and are stratified by year of compilation, communicative function, and pupil's sex. All together total 522 individual files and *c.*93,000 words. The basal readers are also stratified by year of publication and communicative function, adding to 21 files and *c.*79,000 words.

(II) *Data transcription*. The original surveys have been prepared in various formats: PDF digitised images; TXT files with raw text and original spelling;

XML files coded in TEI (Lite), with metadata in individual headers; part-of-speech tagged files (CLAWS); and semantically tagged files (USAS). Word frequency lists are also provided.

(III) *Online interface*. The corpus will be made freely available online, with a user-friendly interface that allows for browsing, searching and downloading search results.

The APU Writing and Reading Corpus 1979-1988 thus presents itself as a new resource tool with replicable methodology and objective empirical evidence that will be of interest to academics and school teachers as well as to school material developers and policy makers.

References

Cambridge Assessment (2016, November). Aspects of Writing in 16+ English examinations between 1980 and 2014. Retrieved from http://www.cambridgeassessment.org.uk/events/aspects-2016/

English in the National Curriculum. (1989). Department of Education and Science and the Welsh Office. London Her Majesty‘s Stationary Office (HMSO).

Foxman, D.D., Hutchinson, D. & Bloomfield, B. (1991). The APU experience: 1977-1990 (Great Britain. Department of Education and Science. Assessment of Performance Unit). [London]: School Examinations and Assessment Council [SEAC EMU].

Grammar for Writing? The impact of contextualised grammar teaching on pupils‘ writing and pupils‘ metalinguistic understanding. Project led by D.A. Myhill and S.M. Jones, University of Exeter. Retrieved from http://education.exeter.ac.uk/research_projects_detail.php?id=18

Hudson, R. & Walmsley, J. (2005). The English patient: English grammar and teaching in the twentieth century. Journal of Linguistics, 41 (3), 593-622.

Ivanic, R., McEnery, T., & Smith, N. (1996-2000). Lancaster Corpus of Children’s Project Writing. Retrieved from http://www.lancs.ac.uk/fass/projects/lever/index.htm

The Oxford Children’s Corpus. See Wild, K., Kilgarriff, A., & Tugwell, D. (2013). The Oxford Children’s Corpus: Using a children’s corpus in lexicography. International Journal of Lexicography, 26 (2), 190-218.

Sealey, A., & Thompson, P. (2004). ‚What do you call the dull words?‘ Primary school children using corpus based approaches to learn about language. English in Education, 38 (1), 80-91.

Sealey, A., & Thompson, P. (2006). ‚Nice things get said‘: Corpus evidence and the National Literacy Strategy. Literacy, 40 (1), 22-8.

• • •

Jiří Zámečník
Albert-Ludwigs-Universität Freiburg
jiri.zamecnik@frequenz.uni-freiburg.de

Mode dependent information density

In his classic paper, Shannon (1948) defines communication as information transmission over a noisy channel with a limited bandwidth. He assumes the speakers to behave rationally and use the channel efficiently, that is transmitting at a rate close to the channel capacity. Striving to explain how the inherent noisiness of the channel might be overcome, Jaeger (2006) suggested his theory of Uniform Information Density, claiming that the signal should also be uniformly smooth – without peaks and troughs in the rate of information transmitted (i.e. information density).

These two theories, however, leave one question open: what limits the channel capacity? Assuming that the channel capacity is limited both at the side of the transmitter and the receiver, I explore whether language producers adjust the information density based on their expectations of the channel capacity. As a first step in this exploration, I built a tool based on Mitchell's (2011) integrated measure (combining a parser, an n-gram model and a semantic model) that allows me to compare the information density of two sets of data, differing only in the mode of their production: spoken vs. written.

In the work-in-progress report, I present the current state of the project. I explain the build-up of the model as well as the test data and present the results yielded by the adopted method. These seem to be counterintuitive: the information density of written texts is lower than of the spoken ones. I discuss the possible reasons for such an observation and propose further options for exploration.

Being able to quantify the difference in acceptable information density between the two modes should prove beneficial for future applications that transform data from one mode to another, i.e. speech recognition tools creating transcripts or text-to-speech synthesizers. Moreover, the methodology used seeks to provide a link between carefully arranged laboratory experiments and a big data approach to language exploration.

References

Jaeger, T. F. (2006). Redundancy and syntactic reduction in spontaneous speech. Ph.D. thesis, Stanford University, Stanford, CA.

Mitchell, J. J. (2011). Composition in distributional models of semantics. Ph.D. thesis, University of Edinburgh, Edinburgh.

Shannon, C. (1948). "A mathematical theory of communications." Bell Systems Technical Journal, 27 (4), 623–656.

• • •

Posters

Sumie Akutsu
Toyo University
akutsu@toyo.jp

Translation Learner Corpus: First Steps Using a Corpus Approach to Teach Mediation Skills

This poster presentation discusses the design and rationale of creating a Translation Learner Corpus using a famous Japanese literary work in conjunction with learners' translations. It illustrates the advantages of using a Japanese literary work as a source text to compile a translation learner corpus and examines the results of preliminary analysis finding common errors among learners' translations in English. Finally, it demonstrates how a translation learner corpus from the translated texts of university students in Japan can cultivate mediation skills between Japanese (L1) and English (L2), thereby bridging the gap between cultures.

The short story, "Run, Melos!", written by the author Osamu Dazai has been used in a university course for translation studies. Using the widely-read short story in secondary education in Japan, the aim is to see if cultural and linguistic differences can be found in the translations in terms of interpretation and language usage of learners of English. A translation learner corpus was compiled during the course with the aim of being analysed like any other learner corpus to identify common errors.

In recent years, translation in the language classroom has been criticized by advocates of the communicative approach to teaching (Cook 2010), which is in spite of the Common European Framework of Reference for Languages defining translation as both an effective means of language learning and as a mediation skill in today's globalized world. As Granger, Lerot, and Petch-Tyson (2003) suggest, a further benefit can be derived from doing more advanced research using a bilingual translation corpus, as corpus-based contrastive linguistics and translation studies can be complementary in terms of research methodology, interests and objectives.

The average Japanese student in university is understandably much more coherent and expressive in Japanese than English; therefore, it is natural for them to struggle to put their advanced Japanese into simple English. To overcome this, it is important to raise awareness of the fact that a language has a culture behind it, and that word-by-word translation between two languages or cultures is not always possible. While students are encouraged to avoid direct translation, they will also learn about the major causes of collocation errors.

Translation activities can have further pedagogical benefits in raising language awareness where learners develop mediation skills between English and Japanese in the process of translating texts and common errors are pinpointed and reified. The poster presentation will demonstrate some examples of these common errors, and how their elucidation then informed the design of subsequent translation activities to highlight the potential errors. The poster concludes students can become autonomous language users with better interpretive translation skills and mediation strategies

through realizing linguistic and cultural differences between Japanese and English languages.

References

Cook, G. (2010). Translation in Language Teaching. Oxford: Oxford University Press.
Dazai, O. (1954). Hashire Merosu [Run, Melos!]. Tokyo: Shinchosha.
Granger, S, Lerot, J. & Petch-Tyson, S. (eds.) (2003). Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Amsterdam: Rodopi.

• • •

Shariya Algama
Justus Liebig University Giessen
shariya.d.algama@anglistik.uni-giessen.de

Here and *men* as Discourse Markers in Spoken Sri Lankan English: A Corpus Based Pilot Study

Discourse markers in varieties of English are known for providing a wealth of insight to discourse strategies and cultural expectations in the area of world Englishes (cf. Valentine 1991: 333; Lange 2009). Indeed, the acculturation of English in 'new' contexts would also mean pragmatic nativisation, taking into consideration the cultural norms and values of societies that find ready expression in indigenous languages. It is important to remember that values and cultural expectations that Western researchers are accustomed to may be communicated differently in a variety of English, but also that the very values a variety of English is moulded to express may be different. Looking at the pragmatic nativisation of a variety of English will reveal interesting linguistic features of the language, as well as norms and expectations of the culture in which the language is situated. Sri Lankan English (SLE) is one such variety that can be studied for elements that are pragmatically relevant as it is a linguistically nativised variety of English (cf. Mukherjee 2012: 198; Bernaisch 2015). This paper studies discourse marker usages of *here* and *men* specific to SLE using the pilot corpus of the spoken component of the International Corpus of English – Sri Lanka (ICE-SL [S75]). The data is compared to that of ICE-Great Britain and ICE-India, the former to isolate the two features in SLE and the latter to rule in or out a possible epicentral influence from Indian English (IndE) as IndE and SLE are both South Asian varieties of English. As for the lexicon of SLE, it would consist of words not in the British English (BrE) lexicon (for example, hybrid compounds) and words that exist in the BrE lexicon, but with truncated, different or added functions. Therefore, in this instance with *here* and *men*, a code copying framework gives insight to a repertoire of words in the SLE lexicon that are words already existing in English's common core shared by varieties, but

with additional, pragmatic functions in SLE. These added functions can be traced to influence from Sinhala and Tamil, two of the indigenous languages used in Sri Lanka and their expression of politeness norms and values particular to interaction in Sri Lanka. Below are two examples for the discourse marking use of *men* and *here* in SLE from the corpus that show an interpersonal function:

Ex:

(1)

<$A><ICE-SL:S1A-003#92:1:A>We have been eating all this time <O>laugh</O>
<$B><><ICE-SL:S1A-003#93:1:B>Yeah *men* I eat slowly <,> when <w>I'm</W> talking <„>
(S1A1-003 tr2 ma1 13-11-29.txt – ICE-SL [S75])

(2)

<$C><#><[>But</[></{> <}><->the</-> <=>the</=></}> thing is I'm enjoying my single life
<$A><O>cough</O>
<$C><#><}><->It's just</-> <=>I'm just</=></}>
<$A><#>*Here* <}><->I would</-> <=>I would</=></}> enjoy it as well
<$C><#><O>cough</O>
<$A><#>If I was <{><[>I would like</[>
(S1A1-002 tr2 ma0 13-11-08.txt- ICE-SL [S75])

Additionally, understanding how exactly a variety of English is used in the articulation of these cultural norms and values also goes towards understanding the extent to which the language is nativised in its 'new' context.

References

Bernaisch, T. (2015). The Lexis and Lexicogrammar of Sri Lankan English. John Benjamins.

Lange, C. (2009). 'Where's the Party yaar!'Discourse Particles in Indian English. World Englishes–Problems, Properties and Prospects. Amsterdam and Philadelphia: John Benjamins, 207-226.

Mukherjee, J. (2012). English in South Asian – ambinormative orientations and the role of corpora: the state of debate in Sri Lanka. In A. Kirkpatrick & R. Sussex (Eds.), English as an International Language in Asia: Implications for Language Education, Dorecht: Springer, 191-208.

Valentine, T. M. (1991): Getting the message across: discourse markers in Indian English, World Englishes 10(3), 325-334.

International Corpus of English: Sri Lanka component [S75], available at: <http://www.uni-giessen.de/anglistik/ling/Staff/mukherjee/>

• • •

Stefan Evert
FAU Erlangen-Nürnberg
stefan.evert@fau.de

Measures of productivity and lexical diversity

Quantitative measures of productivity and lexical diversity – such as the type-token ratio (TTR), Baayen's (1992) productivity index P, or Yule's (1944) K – play a key role in many corpus studies. They are used to assess the degree of morphological productivity (Baayen 1992; Evert & Lüdeling 2005), to estimate the size of an author's vocabulary (Gani 1975; Efron & Thisted 1976), to investigate stylometric differences between writers and settle questions of disputed authorship (Juola 2006), to study diachronic changes in grammar (Bentz et al. 2014), to assess the readability and difficulty level of a text (Graesser et al. 2004), to explore the linguistic correlates of dementia (Garrard et al. 2005; Le et al. 2011), and as a feature in the multivariate analysis of linguistic variation (Biber 1988).

However, virtually all of the approaches and quantitative analyses found in the literature suffer from a number of serious methodological problems:

1. In most cases, no effort is made to assess the uncertainty introduced by sampling variation (Baroni & Evert 2007) and it remains unclear whether the difference between two observed productivity values can be deemed significant. Even if sophisticated statistical LNRE models (Baayen 2001) are used to compute standard errors, authors fail to take the non-randomness of natural language and the variability of estimated model parameters into account.

2. Most quantitative measures depend systematically on sample size (i.e. the size of the corpus or text for which they are computed). This can easily be demonstrated for TTR and P (as argued e.g. by Evert & Lüdeling 2001), but has also been observed with sophisticated LNRE models (Baayen 2001, Fig. 5.12 on p. 182). Normalizing all texts to the same length is neither a practicable nor a satisfactory solution.

3. Measures are highly sensitive to the presence of a small number of (lexicalized) high-frequency types, which can have a stronger influence on productivity values than the richness of productively formed types. Measures are also sensitive to noise introduced e.g. by typographical mistakes, OCR errors and lack of spelling normalization. Different editorial conventions as well as choices made by the researcher (e.g. whether to consider lowercase and uppercase spellings as distinct types) can lead to further spurious differences in lexical diversity.

4. Many quantitative measures do not have a clear and obvious linguistic interpretation. This holds in particular for the more sophisticated approaches (such as LNRE models) that account for some of the aforementioned problems; but even for simpler measures, their relation to intuitive notions of productivity and lexical richness remains unclear.

My poster gives an overview of established measures of productivity and lexical richness and discusses the four methodological problems listed above in detail. I will also suggest improved approaches and quantitative measures. All findings will be

illustrated with simulation experiments as well as a case study based on a large collection of English literary texts (Capitanu et al. 2016).

References

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. Yearbook of Morphology 1991, 109–149.

Baayen, R. H. (2001). Word Frequency Distributions. Dordrecht: Kluwer Academic Publishers.

Baroni, M. & Evert, S. (2007). Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. In Proceedings of ACL 2007, 904–911. Prague, Czech Republic.

Bentz, C., Kiela, D., Hill, F., Buttery, P. (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. Corpus Linguistics and Linguistic Theory, 10 (2), 175–211.

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Capitanu, B., Underwood, T., Organisciak, P., Cole, T., Sarol, M. J. & Downie, J. S. (2016). The HathiTrust Research Center extracted feature dataset (1.0). http://dx.doi.org/10.13012/J8X63JT3.

Efron, B. & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? Biometrika, 63 (3), 435–447.

Evert, S. & Lüdeling, A. (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? In Proceedings of Corpus Linguistics 2001, 167–175. Lancaster, UK.

Gani, J. (1975). Some stochastic models in linguistic analysis. Advances in Applied Probability, 7 (2), 232–234.

Garrard, P., Maloney, L. M., Hodges, J. R. & Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain, 128 (2), 250–260.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36 (2), 193–202.

Juola, P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval, 1 (3), 233–334.

Le, Z., Lancashire, I., Hirst, G. & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Literary and Linguistic Computing, 26 (4), 435–461.

Lüdeling, A. & Evert, S. (2005). The emergence of productive non-medical -itis. Corpus evidence and qualitative analysis. In S. Kepser & M. Reis (eds.), Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives. Berlin: Mouton de Gruyter, 351–370.

Yule, G. U. (1944). The Statistical Study of Literary Vocabulary. Cambridge: Cambridge University Press.

• • •

Yu-Hua Chen
University of Nottingham Ningbo China
Yu-Hua.Chen@nottingham.edu.cn

Simon Harrison
University of Nottingham Ningbo China
Simon.Harrison@nottingham.edu.cn

Michael Stevens
University of Nottingham Ningbo China
Michael-Paul.STEVENS@nottingham.edu.cn

Beyond Borders, Beyond Words: A New Multimodal Corpus of L2 Academic English from A Sino-British University

The Corpus of Chinese Academic Written and Spoken English (CAWSE) is an ongoing project which aims to build a large collection of Chinese students' English language samples from one of the few English-medium instruction (EMI) universities in China. The campus creates a unique environment for teaching and learning and also provides exciting opportunities for linguistic studies into Academic English from diverse theoretical and analytical perspectives. The project collects students' language samples from a variety of assessment tasks (both written and spoken) and speech events (spoken and multi-modal) from the campus. The final product of CAWSE will offer open-access electronic resources (including a multi-modal subcorpus) available for all researchers and practitioners who are interested in a wide range of topics, including for example Second Language Acquisition, English for Academic Purposes, World Englishes, and many other aspects of the Written and Spoken English unique to this new corpus.

The spoken and multimodal subcorpora are expected to reach no less than one million tokens and no less than 100 speech events. While the majority of the corpus will be text-based, including the orthographically transcribed spoken subcorpus, one of the greatest challenges is the daunting task of transcribing L2 speech data. A pilot corpus including oral presentation (scripted speech) and group discussion (spontaneous speech) have been transcribed first using state-of-the-art technology, i.e. speech recognition, and then manually checked and edited. While an increasing amount of speech data is currently being collected, the approach of adopting ‚crowdsourcing' (i.e. providing audio/video recordings for interested researchers and students for their own research and receiving completed transcription data in return) will also be adopted. This paper will introduce this unique CAWSE corpus and then discuss the challenges and issues of using various innovative approaches in constructing a L2 multimodal corpus.

• • •

Bridget Jankowski
University of Toronto
bljankowski@gmail.com

There's no place like this place, any place! Variable quantified compounds of location in Ontario English

This study presents a quantitative sociolinguistic analysis of the English quantified compounds of location in a large corpus of sociolinguistic interviews from thirteen communities in Ontario, Canada. *Some–, any–, every–* and *no–* combine with *–where* and *–place* to form four doublets, in (1)–(4), which can serve as adverbs or nouns:

(1) *Somewhere* in the house we used to have pictures […] I have a picture *some-place* in the house. (BH, male, b. 1939, KL)
(2) a. There's nothing, no cars, no objects *anywhere.* (CC, female, b. 1983, TOR)
b. There's no lights *anyplace.* (KR, male, b. 1925, BEA)
c. I haven't found *any place* I'd rather be. (EB, female, b. 1930, HB)
(3) a. There's like eleven of us traveling *everywhere.* (JF, male, b. 1963, SP)
b. There's people *everyplace* there nowadays. (SB, female, b. 1941, W-BB)
c. *Every place* we stopped at, we ate some. (MS, male, b. 1991, SP)
(4) a. I had *nowhere* to ride my four-wheeler. (RC, male, b. 1979, KL)
b. I had *no place* to eat. (AS, male, b. 1963, TOR)

The communities examined contrast by population size and urban vs. rural, with data from over 588 individuals born 1912–2000, permitting investigation across time, geography, and social factors (sex, occupation, education level).

The *–where* variants are older. *Nowhere* is found in Old English, with the others attested in Middle English (*somewhere/everywhere* c1200–1225, *anywhere* ~1400). First attestations for the *–place* variants range from 1819 (*anyplace*) to 1928 (*everyplace*) (OED Online). The *–place* forms are considered "colloquial N. American" (Quirk et al. 1985: 782), and "*dial.* and *U.S.*" (OED online), indicating potential correlations with social factors.

Unlike the related pronominal quantifiers with *–body* and *–one* (see, e.g. D'Arcy et al. 2013), the compound locatives have not received in-depth examination. Report of the *any–* forms appears (13% *–place*) in only one 20th century North American dialect atlas collected under the Linguistic Atlas Project (http://www.lap.uga.edu/). The Dictionary of American Regional English similarly contains an entry only for *anyplace* (Cassidy 1985: 73).

In Ontario English, the *–where* variant predominates (87% overall, N=2,394), but shifts from 72% in individuals born before 1950 to 95% in those born after 1950. Quantifier *no–* stands out as having the highest rate of *–place* forms (11%), even in the youngest individuals. In fixed-effects logistic regression, year of birth (pre- vs. post 1950) is significant. There are also effects of community and education: larger urban centres and individuals with more education use mostly *–where.*

These results can be interpreted under the cascade model (Trudgill 1972; Labov 2007), which predicts rural communities further from large population centres will retain declining forms. Interestingly, although *–place* compounds are relatively recent innovations, they never fully penetrated the vernacular and are in now in steep decline, possibly due to increased levels of formal education in speakers born after 1950. This study shines light on an understudied North American colloquialism, and how a vernacular form winds its way in to — and possibly out of — the grammar.

References

Cassidy, F. C. (1985). Dictionary of American Regional English, Vol. 1. Cambridge, MA: Harvard University Press.

D'Arcy, A., Haddican, B., Richards, H., Tagliamonte, S. A. & Taylor, A. (2014). Asymmetrical trajectories: The past and present of -body/-one. Language Variation and Change 25(3): 287-310.

Labov, W. (2007). Transmission and Diffusion. Language 83(2), 344-387.

Kretzschmar, Jr., W. A., (Ed.). "anywhere". Linguistic Atlas Project (LAP). (Online). Accessed 13 December 2016: http://www.lap.uga.edu/LICHEN/html/gui/index.php?filt_question=anywhere &submitSearch=Search&aspect=question# .

OED Online. December 2016. Oxford University Press.

Quirk, R., Greenbaum, S., Leech, G. et al. (1985). A Comprehensive Grammar of the English Language. New York: Longman.

Trudgill, P. J. (1972). Linguistic change and diffusion. Language in society 3(2), 229-252.

• • •

Jakub Jehlička
Charles University
jakub.jehlicka@ff.cuni.cz

Eva Lehečková
Charles University
eva.leheckova@ff.cuni.cz

Towards multimodal constructions: bounded events, bounded gestures

In our poster, we focus on a broad research question of how the way verbs structure events corresponds with formal properties of co-speech gestures in English spontaneous spoken production. In particular, we examine whether the semantic feature of boundedness expressed on verbs correlates with the boundedness marked by co-speech gestures.

The bounded/unbounded distinction is one of the lexicosemantic features that have been traditionally associated with verbal semantics in the aspectuality studies that focus on lexicosemantic classification of events (cf. Vendler 1967, Smith 1997, Filip 1999 among others). More specifically, there are theoretical approaches that distinguish boundedness from aspect as well as telicity (cf. Depraetere 1995), expecting boundedness to refer to events delimited temporally, namely happening within a certain time span. Recently, Croft (2012) has proposed a model of linguistic boundedness that delimits two kinds of bounded events: a) t-bounded events which are temporally framed (e.g. *Yesterday, I took a bus*), and b) q-bounded events that are qualitative state bounded (i.e. they have an inherent endpoint, e.g. *I read a book* [q-boundedness thus equals to the notion of telicity]).

Relation of co-speech gestures and eventuality has been examined in a number of studies on English (McNeill & Levy 1982, Becker et al. 2011, Parrill et al. 2013) and across languages (Duncan 2002). The research has so far revealed systematic correspondences between semantics of verbs and formal features of accompanying gestures, such as shorter and less complex gestures used in achievements when compared with activities or unbounded gestures associated significantly more often with progressives.

In our study, we present results of an analysis of an approximately 90-minute sample of English spontaneous spoken interactions obtained from a multimodal corpus (Carletta 2006). The sample consists of a series of interactions of 3-4 native speakers captured during business meetings (10 English speakers in total). In line with the previous research, we coded the verbs in the sample for aspect and q/t-boundedness. The gestures were coded for boundedness (Boutet 2010, Cienki 2016). The annotation was performed by two independent coders and the inter-annotator agreement was measured (with Cohen's $\kappa>0.85$). First, association of linguistic and gestural boundedness was computed in R (R core team, 2016) (applying mixed-effects models in order to control the inter-speaker variation). Additionally, the effect of aspect was measured.

As for the main results, we have observed a significant tendency of the qualitative state bounded verbs and gestures with accentuated point of movement (i.e. bounded) to co-occur whereas in t-boundedness, no similar association was attested. We assume that these results are due to the fact that t-boundedness is usually expressed lexically whereas q-boundedness in English is formally more opaque, and therefore more prone to be marked explicitly by gesture.

References

Becker, R., et al. (2011). Aktionsarten, speech and gesture. Proceedings of the GESPIN 2011.

Carletta, J. (2006). Announcing the AMI Meeting Corpus. The ELRA Newsletter, 11(1), 3–5.

Boutet, D. (2010). Structuration physiologique de la gestuelle: modèle et tests. LIDIL, 42, 77–96.

Cienki, A. et al. (2016). Linguistic aspect and tense and gestural movement quality in French, German, and Russian utterances. Talk at ISGS Conference 2016, Paris, July 2016.
Croft, W. (2012). Verbs: aspect and causal structure. Oxford & New York, NY: Oxford University Press.
Depraetere, I. (1995). On the necessity of distinguishing between (un)boundedness and (a)telicity. Linguistics and Philosophy, 18, 1–19.
Duncan, S. D. (2002). Gesture, verb aspect, and the nature of iconic imagery in natural discourse. Gesture 2(2), 183–206.
Filip, H. (1999). Aspect, eventuality types, and nominal reference. New York, NY: Garland Pub.
McNeill, D., & Levy, E. T. (1982). Conceptual Representations in Language Activity and Gesture. In R. J. Jarvella, & W. Klein (Eds.), Speech, Place, and Action. Studies in Deixis and Related Topics. Chichester, NJ: John Wiley & Sons, 271–295
Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures?: Insights from crosslinguistic variations and similarities. Gesture, 5(1–2), 219–240.
Parrill, F., Bergen, B. K., & Lichtenstein, P. V. (2013). Grammatical aspect, gesture, and conceptualization: Using co-speech gesture to reveal event representations. Cognitive Linguistics, 24(1), 135–158.
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Vendler, Z. (1967). Linguistics in philosophy. Ithaca, NY: CUP.

• • •

Joonas Kesäniemi
Helsinki University Library
joonas.kesaniemi@helsinki.fi

Turo Vartiainen
University of Helsinki
turo.vartiainen@helsinki.fi

Tanja Säily
University of Helsinki
tanja.saily@helsinki.fi

Agata Dominowska
University of Helsinki
agata.dominowska@helsinki.fi

Aatu Liimatta
University of Helsinki
aatu.liimatta@helsinki.fi

Terttu Nevalainen
University of Helsinki
terttu.nevalainen@helsinki.fi

Making new use of old research: the Language Change Database

Thirty years ago the digital turn in linguistics changed the empirical basis for studies of language change. Since the publication of the first corpora, such as the Helsinki Corpus of English Texts, diachronic corpora have multiplied and grown in size. While the material basis of the research in the field has gradually become more unified and accessible, this is not necessarily the case with publications: English historical linguistics has a long publishing tradition in monographs and articles in collective volumes, which means that the fast-growing empirical work on the changing English language, and the data on which the research are based, is often scattered and difficult to retrieve.

In 2014, we launched a project to remedy this situation. Our aim is to make research more accessible and cumulative by compiling an online Language Change Database (LCD), which will be made freely available to the research community on a wiki-style platform which the researchers can update themselves (Nevalainen et al. 2016). The LCD, which will be published in 2018, draws together a growing collection of empirical studies on the history of English. Each LCD entry is annotated for several features according to which the database can be queried, including grammatical and sociolinguistic keywords, the periods studied, and bibliographical details. The

numerical data discussed in the articles are included as Excel files, which the end users can download and reanalyse on their own computers. Each entry also includes an abstract and a summary of the main findings of the study.

We envisage a wide variety of uses for the LCD in both research and teaching. The database gives an easy access to work published on different topics, periods and data sources, thus providing versatile baseline data for future research and replication of earlier findings with new data sets. As the database grows, it becomes possible to test assumptions on issues such as the rate and direction of language change in particular grammatical domains, periods and registers. It will also be possible to carry out meta-analyses on the numerical data included in the LCD by applying a variety of computational techniques (Baxter & Croft 2016). This approach offers new perspectives, for example, on a given period and invites further computational modelling of the diffusion of linguistic change.

In our poster, we focus specifically on the most recent developments of the database, including a normalisation tool that can be used both to validate the data in the articles (i.e. to check which version of the corpus has been used as the baseline for normalised frequencies) and also to normalise the raw frequencies reported in the articles. For this purpose, we have described the numerical data in the entries using RDF (Resource Description Framework) Data Cube vocabulary, which allows for flexible data combination across data sets and greatly improves the machine-readability and reuse of research data.

We welcome feedback and suggestions from the ICAME community on both theoretical and practical questions related to the database.

References

Baxter, G. & Croft, W. (2016). Modelling language change across the lifespan: Individual trajectories in community change. Language Variation and Change, 28(2), 129–173.

Nevalainen, T., Vartiainen, T., Säily, T., Kesäniemi, J., Dominowska, A. & Öhman, E. (2016). Language change database: A new online resource. ICAME Journal, 40(1), 77–94.

• • •

Ilka Mindt
Paderborn University
mindt@mail.uni-paderborn.de

The grammar and semantics of the adjective *chosen*

The word *chosen* typically occurs as a verb as in (1). However, it might be used as an adjective in attributive position as well as demonstrated in (2).

(1) *For all of the statistical analyses, an alpha level of.05 was* **chosen** *to indicate a significant difference.* (COCA 2012 ACAD LanguageSpeech)
(2) *Her grandfather always said that God gave David's words and music the power to lift hearts and minds from petty differences to God's majesty and the blessings He had poured upon His* **chosen** *people.* (COHA FICTION 2001 Unspoken)

It is interesting to note that the adjective *chosen* is not listed as such in the current learner dictionaries (OALD, LDOCE). LDOCE only gives the phrase *the chosen few*.

The assumption in the present project is that the verbal use of *chosen* is canonical whereas the adjectival use shows non-canonical features. The main aim of this investigation is in how far the adjective *chosen* has canonical or non-canonical features of adjectives. The definition of ‚canonical' and ‚non-canonical' as put forward by Huddleston and Pullum (2002) will be elaborated. Rather than equating ‚canonical' with „basic […] constructions" (2002: 46) or information packaging (compare 2002: 1365ff), I will refer to parts of speech and their grammatical and lexical uses. Canonicalness will be expressed in terms of frequency distributions, where canonical uses are more frequent as opposed to non-canonical ones. Quantitative data distributions in terms of grammar, the co-text, spelling alternatives (i.e. whether *chosen* is spelt with an initial capital letter or not) as well as lexical meanings will be analysed. In particular, this study will focus on lexical and grammatical analyses from a diachronic point of view. The quantitative and qualitative analysis will be based on data taken from COHA and COCA.

References

Huddleston, R. & Pullum, G. (2002). The Cambridge Grammar of the English Language. Cambridge: Cambridge University Press.
[LDOCE] (52009). Longman Dictionary of Contemporary English. Harlow: Pearson.
[OALD] (92015). Oxford Advanced Learner's Dictionary. Oxford: Oxford University Press.

• • •

Aika Miura
Tokyo University of Agriculture
dawn1110am@gmail.com

The challenges with annotating requestive speech acts in learner corpora

This paper reports the challenges associated with the use of spoken corpora in investigating the pragmatic competences of learners of English as Foreign Language. The author also presents how she overcame the difficulties in building multi-layered annotation schemes of requestive speech acts, using the UAM CorpusTool (O'Donnel, 2013).

She examined the learners' requests in shopping role plays with interlocutors in the National Institute of Information and Communications Technology Japanese Learner English Corpus. Drawing on the coding scheme developed by Blum-Kulka et al. (1989), a manual annotation scheme was constructed to identify and classify requests into different strategies, referring to pragmalinguistic features. For example, *I want to try on this jacket* is a direct strategy using the desire verb *want*, and *Can I try on this jacket?* is a conventionally indirect strategy using the modal *can*. Situations where each speech act occurred were also identified for clarifying the functions of requests, such as 'asking for permission to test an item at a shop'.

However, as the learners' spoken data may be interactionally incomplete, socially inappropriate, and grammatically erroneous, building additional multi-layered annotations to overcome the following challenges was necessary.

The first challenge is characteristic to spoken data, which contain dysfluency and raise issues of segmentation. In role plays, learners' utterances may be interrupted by the interlocutor so that one single unit of a speech act can be divided into two non-adjacent parts. Therefore, a single unit was classified into a main and subordinating segment, of which the latter was a remaining utterance of the former. Thus, 'repair' in interactions was also identified (Kasper & Ross 2007) to examine the ratio of redundant features, by classifying the learner production into 'rephrasing', 'repeating one's own utterances', or 'echoing the interlocutor's utterances', based on extralinguistic tags originally annotated to the Corpus, as well as referring to the interactional information.

The second challenge lies in the learners' low proficiency: some of their utterances lacked not only grammatical correctness but also logic and coherence in given interactions. Therefore, in terms of 'grammatical accuracy/discoursal acceptability', all units were classified into either 'high' or 'low'. The 'high' utterances were the ones which were highly grammatical as well as highly acceptable in terms of discourse (i.e., coherently responding to the interlocutor). On the other hand, the 'low' ones contained some problematic grammatical/discoursal features, and were further divided into 'coherent', 'incoherent', having a 'topic-comment-structure', and the use of 'Japanese'. The 'coherent' ones, although showing some degree of ungrammaticality, were coherent in terms of discourse, but 'incoherent' ones were not. The units iden-

tified as 'topic-comment-structure' had an ungrammatical topic-comment structure influenced by the mother tongue, such as "And its color is black".

The final challenge remains unresolved as it is related to the issue of annotating pragmatic appropriateness to the Corpus, which seems unfeasible. The judgement test was conducted to examine whether judges could agree on assessing the learners' sociopragmatic competences. Frequently observed linguistic patterns with different strategies in frequently occurring situations were extracted verbatim from the Corpus, and 20 English-language instructors of tertiary education in Japan (10 native speakers of English and 10 of Japanese) evaluated them for the degree of appropriateness: appropriate (i.e., sufficiently polite), a little appropriate (i.e., a little too polite or little impolite), or inappropriate (i.e., too polite or impolite). Thus, 10 different linguistic features 'asking for exchanges or return of items', 6 requests 'asking for permission to test an item', and 8 features 'expressing intention to buy an item' were investigated. The first situation showed the highest Kendall's Coefficient of Concordance, *W* (0.64), but the scores in other two situations were not significantly high.

References

Blum-Kulka, S., House, J. & Kasper, G. (1989). Cross-cultural Pragmatics: Requests and Apologies. Norwood, NJ: Ablex.

Kasper, G. & Ross, S. (2007). Multiple questions in oral proficiency interviews. Journal of Pragmatics 39, 2045-2070.

O'Donnell, M. (2013). UAM CorpusTool (Version 3.0) [Text annotation tool]. Retrieved from http://www.corpustool.com/download.html

• • •

Magnus Nissel
University of Giessen
magnus.nissel@anglistik.uni-giessen.de

***High motor & deceptive athleticism*: An analysis of basketball discourse in tweets, comments, and blog posts**

This pilot study is the start of a project investigating the stylistics and discourse patterns of sports talk on the Internet using a large-scale quantitative approach. The research presented in this poster focuses on keywords and -phrases containing adjectives and adverbs, with the goal of adding empirical insights to recent discussions of racial biases and stereotypes in the description of NBA athletes.

Within sports journalism, observers have noted that there appears to be an inventory of labels – such as *good fundamentals*, *high motor*, *cerebral*, and *sneaky athletic* – that is disproportionately applied to white players (cf. Yoder 2014). In a similar vein,

an analysis of 565 scouting reports has shown that young players are predominantly compared to established players of similar skin color (cf. Garcia 2014). During the 2015-16 season, NBA star Kobe Bryant told an African American teammate to *stop going to the hole like a light-skinned dude* (cf. Schilling 2016) and play more like a *dark-skinned dude* (ibid.).

The use of a (monitor) corpus of different text types, from 140-character messages to longform articles, enables a nuanced characterization across online platforms. An automation has been set up to collect posts on two social media sites (Twitter and Reddit) as well as blog articles for each day of the current season of the National Basketball Association. For this poster, the time period between October 25th, the first gameday of the regular season, and February 16th, the last gameday before a mid-season break, has been selected. The table below provides an overview of the distribution of words in the sample corpus.

Words by month and text source

	2016-10	2016-11	2016-12	2017-01	2017-02	All
Blogs	495,961	3,306,493	3,128,492	3,163,617	2,407,535	12,502,098
Reddit	4,027,772	17,131,868	19,430,708	25,092,903	13,684,971	79,368,222
Twitter	803,113	4,304,630	4,999,401	5,210,763	1,756,301	17,074,208
All	5,326,846	24,742,991	27,558,601	33,467,283	17,848,807	108,944,528

While all blog articles and almost all collected Reddit discussion threads relate to basketball, the Twitter component offers no straightforward way to distinguish posts about the NBA from any other topic. For that reason, a wordlist consisting of basketball terms as well as team and player names is used to narrow down the tweets for further analysis.

In addition to traditional keyword analysis, methods of topic modelling are employed in order to investigate where - and to what degree - stereotypical descriptions occur in the discourse and whether there are indeed certain clusters that can be explained by racial or cultural biases.

References

Garcia, W. (2014). The Black-White Thing: Racial Biases in NBA Scouting Comparisons. Retrieved from http://hardwoodparoxysm.com/2014/09/02/black-white-racial-biases-nba-scouting-comparisons

Yoder, M. (2014). Your white guy code word power rankings. Retrieved from http://awfulannouncing.com/2014/your-white-guy-code-word-power-rankings.html

Schilling, D. (2016). Kobe Bryant‘s ‚light-skinned‘ remark hints at NBA‘s peculiar racial politics. The Guardian. Retrieved from https://www.theguardian.com/sport/2016/jan/06/kobe-bryant-steph-curry-light-skinned-remark-hints-at-nbas-peculiar-racial-politics

• • •

Sara Norja
University of Turku
skmnor@utu.fi

The Mirror of Alchemy: Alchemical texts as a source for corpora

The source material for historical corpora – usually scholarly editions – are often not entirely accurate with regard to the original text, sometimes even if the editions in question claim to be faithful representations of the original. Documentary editing of historical material can provide a solution to that issue: the documentary method seeks to provide a faithful witness of the original material and to avoid any ahistorical, editorial insertions. The problem even with accurate documentary editions, of course, is that if they are in print, a corpus compiler will still have to digitise the material; and with any manipulation of the original text, the risk of error grows greater. Thus, digital documentary editions are of the most use to historical corpus linguists (Lass 2004).

Alchemy was one of the first experimental sciences in the Middle Ages and influenced the development of chemistry. A multitude of English medieval alchemical manuscript texts survive, written in both Latin and the vernacular. However, the uncharted material vastly outnumbers the texts edited so far (Grund 2013). Vast amounts of early modern alchemical manuscript material also exist and remain largely unedited. In order for corpus linguists to utilise this branch of early scientific writing, more alchemical texts need to be edited – preferably in a digital form immediately compatible with corpus search tools. This poster will present one such editorial project.

The Mirror of Alchemy (*MoA*) is a well-known alchemical work, previously (and spuriously) attributed to Roger Bacon (c. 1214–1292). My doctoral research concerns the seven extant, as yet unexplored English-language manuscript versions of *MoA* (15th–17th centuries). A TEI XML based, open-access digital scholarly edition (DSE) of the versions of *MoA* will form part of my PhD dissertation: the DSE will be documentary, but combined with a reader-friendly best-text edition. My editorial principles build on the standards for the digital editing of Middle English texts proposed by Marttila (2014). Marttila edited Middle English culinary recipes; I will adapt his methods to suit the editing of alchemical texts and to accommodate the early modern versions of *MoA*. The TEI XML encoding will make the DSE searchable and usable for quantitative research. The edition aims to be an accurate representation of the manuscript texts, and will encode aspects of the manuscript text including rubrication, changes in hand, marginal notes and code-switching. The aim is to also include POS tagging.

Early English scientific writing has long been an important part of corpus compilation (e.g. the *Middle English Medical Texts* and *Early Modern English Medical Texts* corpora). Interpreting the world of early science through an eventual corpus of alchemical texts will only be possible with more – and accurate – editions of alchemical material (such as e.g. the edition of Isaac Newton's alchemical writings, Newman 2005). My edition of *MoA* will contribute towards this goal.

References

Grund, P. J. (2013). Editing alchemical texts in Middle English: The final frontier?. In V. Gillespie & A. Hudson (Eds.), Probable Truth: Editing Medieval Texts from Britain in the Twenty-First Century. Turnhout, Belgium: Brepols, 427–442.

Lass, R. (2004). Ut custodiant litteras: Editions, corpora and witnesshood. In M. Dossena & R. Lass (Eds.), Methods and Data in English Historical Dialectology. Bern: Peter Lang, 21–48.

Marttila, V. (2014). Creating Digital Editions for Corpus Linguistics: The Case of Potage Dyvers, A Family of Six Middle English Recipe Collections. PhD dissertation. University of Helsinki, Department of Modern Languages. [http://urn.fi/URN:ISBN:978-951-51-0060-3, accessed 13 March 2017]

Newman, W. R. (Ed.) (2005). The Chymistry of Isaac Newton. Indiana University. [http://webapp1.dlib.indiana.edu/newton/, accessed 13 March 2017]

• • •

Päivi Pahta
University of Tampere
paivi.pahta@uta.fi

Minna Palander-Collin
University of Helsinki
minna.palander-collin@helsinki.fi

Minna Nevala
University of Tampere
minna.nevala@uta.fi

Arja Nurmi
University of Tampere
arja.nurmi@uta.fi

Turo Hiltunen
University of Helsinki
turo.hiltunen@helsinki.fi

Jukka Tyrkkö
Linnaeus University
jukka.tyrkko@lnu.se

Democratisation and language practices: Introducing the DEMLANG project

Sociocultural processes related to the notion of **democratization** are known to have had an impact on how English has changed particularly during the latter half of the twentieth century. Earlier studies on 20th-century English have documented how features of colloquial and private language are spreading increasingly to the public sphere, and how linguistic systems shift with changing ideologies (e.g. Mair 2006, Leech et al. 2009, Farrelly & Seoane 2012). What is less clear, however, is how this process has influenced language use in a longer diachronic perspective. This applies also to the related, but separate process of **mediatization**, understood as comprising the rapid changes in mass media and the overall increase of mediated communication (e.g. Krotz 2008). This poster introduces *Democratization, Mediatization and Language Practices in Britain, 1700–1950* (DEMLANG), a newly launched research project at the universities of Tampere and Helsinki, Finland, funded by the Academy of Finland 2016–2020. Using large digital data sets, DEMLANG aims to produce new information to describe the two-way relationship of language practices and the sociocultural processes of democratization and mediatization.

The poster provides an overview of the different sub-projects and the ways they plan to make use of a variety of public texts mediating ideologies and values, such

as newspaper texts, political speeches, parliamentary records, and novels. The data comes from corpora and databases of different sizes including Hansard Online (https://hansard.parliament.uk/) and the British Newspaper Archive (http://www.britishnewspaperarchive.co.uk/). New corpora will also be collected, such as The Punch Corpus. These texts will be interrogated through quantitative and qualitative methods to foreground different aspects in the interrelationship of **democratisation** and **mediatisation** and their role in developing and promoting linguistic practices. The phenomena in focus include changes in the rhetoric of expert discourses, explicit linguistic markers of power, stance, voice, identity construction and categorisation, as well as multilingual practices.

References

Farrelly, M. & Seoane, E. (2012). Democratization. In T. Nevalainen & E. C. Traugott (Eds.), The Oxford Handbook of the History of English. Oxford: OUP, 392–401.

Krotz, F. (2008). Media connectivity: Concepts, conditions, and consequences. In A. Hepp, F. Krotz & S. Moores (Eds.), Network, Connectivity and Flow: Key Concepts for Media and Cultural Studies. New York: Hampton Press, 13–31.

Leech, G., Hundt, M., Mair, C. & Smith, N. 2009. Change in Contemporary English: A Grammatical Study. Cambridge: Cambridge University Press..

Mair, C. (2006). Twentieth-Century English: History, Variation, and Standardization. Cambridge: Cambridge University Press.

• • •

Sirkku Ruokkeinen
University of Turku
siruok@utu.fi

Presenting Knowledge of the World: An Analysis of Appraisal in the English Renaissance

When John Florio (1553–1625) presented his translation of the French explorer Jacques Cartier's *Voyages* (1580) to his English readers, he did so wishing to give "no smal commoditie and benefite to this our Countrie of Englande" through the access to the "infinite treasures" of geography and navigation already known to the Spanish, the Portuguese, and the Venetians. The translation, he states, reveals secrets previously unknown to the English seamen and is hence beneficial to the nation as a whole. Similar evaluations stressing the usefulness of geographical, navigational, or historical accounts were a staple of renaissance translators' prefatory matter.

In my poster, I analyze the ways in which works providing knowledge of the world are presented to new audiences within renaissance England. I conduct a computer

assisted analysis of evaluative language within a 40,000 word corpus of translator's prologues and dedications in works of history, geography, and navigation, translated and printed during the sixteenth century.

I utilize the **appraisal framework** (AF) in annotating the tokens of evaluative language. Developed by the Australian systemic functionalists, AF is a model for the categorization and analysis of linguistic resources of emotion and opinion (Martin and White 2005; Martin 2000). Rather than viewing evaluation simply as the speaker's way of construing his or her experience of the world, AF considers evaluation as an interpersonal tool, a method for building solidarity within communities by expressing and upholding communal values. Hence the tokens of evaluation are annotated not only according to their position within the typology of appraisal, but also by their emoter, target, and complexity.

Within the context of renaissance prologue and dedication, appraisal analysis reveals the mix of textual traditions and new ideas and knowledge. For example, while the ancient tradition of humility discourse (see e.g. Curtius 1990; Janson 1964) often manifested as negative evaluations, a more positively toned prosody is also readily apparent: the values of **novelty**, **necessity**, and **usefulness** of the translated content are juxtaposed with the **rudeness** of its new form.

This poster is a part of my PhD dissertation project, in which I study the evaluation of the book in sixteenth-century England, within a 90,000 word corpus of translators' prologues and dedications.

References

Cartier, J. (1580). A Shorte and Briefe Narration of the Two Nauigations and Discoueries to the Northweast Partes Called Newe Fraunce. London. Text Creation Partnership digital edition. Early English Books Online. Web. 27 January 2017.

Curtius, E. R. (1990). European Literature and the Latin Middle Ages. Princeton: Princeton University Press.

Janson, T. (1964). Latin Prose Prefaces: Studies in Literary Conventions. Stockholm: University of Stockholm.

Martin, J. R. (2000). Beyond Exchange: Appraisal Systems in English. In S. Hunston & G. Thompson (Eds.), Evaluation in Text: Authorial Stance and the Construction of Discourse. Oxford: Oxford University Press, 142–75.

Martin, J. R. & White, P. R. R. (2005). The Language of Evaluation: Appraisal in English. Basingstoke: Palgrave Macmillan.

• • •

Christina Sanchez-Stockhammer
LMU Munich
christina.sanchez@fau.de

Creating a mosaic of English language usage with student-compiled micro-corpora

This poster suggests a new way of combining corpus linguistic research and teaching that simultaneously benefits the description of the English language and the development of students' research skills.

If we wish to interpret the world through corpora, as suggested by the theme of this ICAME conference, the corpora used need to reflect the world in a comprehensive manner. Since "the" English language is an abstraction over immensely diverse communicative situations, recent research has placed a strong focus on the corpus-linguistic investigation of variation. For instance, regional variation can be compared with a growing number of ICE Corpora, and ICLE permits the comparison of variation between learners of English with different linguistic backgrounds. However, even corpora for specific subsets of English language usage that do not follow identical collection and coding guidelines still contribute to the comprehensive linguistic description of the language. This applies particularly where situation-specific usage in the sense of Biber & Conrad's (2009) register variation is investigated, and where important differences between the communicative situations may prevent completely shared guidelines. Recent attempts to foster the description of so-far underresearched registers (e.g. Schubert & Sanchez-Stockhammer 2016) have resulted in corpus-based analyses of e.g. hip-hop lyrics (Kreyer 2016) or the language of crossword puzzles (Pham 2016). Each new corpus compiled for a specific register of the English language increases our means of learning more about this particular register, the English language as a whole and variation in general.

In the course of their linguistic studies at university, numerous students write coursework on empirical subjects. For some of these projects, students compile their own corpora in order to answer their research questions (e.g. whether male and female tabloid talk show hosts differ regarding their interruptive behaviour). These early-stage researchers invest a large amount of time and effort to compile corpora on highly specialised topics such as basketball commentaries or the phonetic representation of Russian accent in Hollywood films. However, after the marking of the students' coursework, such corpora are usually not put to any other use and are thus lost to the linguistic research community. This is where my project sets in.

I am currently compiling an open-ended collection of student micro-corpora which will be made available through the University of Erlangen-Nuremberg's online linguistics platform *Erlingo* in early 2017. Users will be able to download the corpus files in text format, accompanied by pdfs with information for each corpus.

My poster gives an overview of the project's goals, its legal background (e.g. the copyright filter), the currently included corpora (with samples) and the types of metadata coded in the corpus files. The poster argues that the publication of student micro-

corpora from high-quality coursework offers important potential for register research (e.g. by laying the foundations for clusters of similar corpora) and for teaching (e.g. by increasing students' motivation and accuracy due to the incentive of product orientation).

References

Biber, D. & Conrad, S. (2009). Register, Genre, and Style. Cambridge: Cambridge University Press.

Erlingo (Erlanger Linguistik Online). http://www.erlanger-linguistik-online.uni-erlangen.de/

ICE (International Corpus of English). http://ice-corpora.net

ICLE (International Corpus of Learner English). http://www.uclouvain.be/en-cecl-icle.html

Kreyer, R. (2016). "Now niggas talk a lotta Bad Boy shit": The register hip-hop from a corpus-linguistic perspective. In C. Schubert & C. Sanchez-Stockhammer (Eds.), Variational Text Linguistics: Revisiting Register in English. Berlin: de Gruyter Mouton, 87-109.

Pham, T. (2016). The register of English crossword puzzles: Studies in intertextuality. In C. Schubert & C. Sanchez-Stockhammer (Eds.), Variational Text Linguistics: Revisiting Register in English. Berlin: de Gruyter Mouton, 111-136.

Schubert, C. & Sanchez-Stockhammer, C. (Eds.) (2016). Variational text linguistics: Revisiting register in English. Berlin: de Gruyter Mouton.

• • •

Yuesen Yang
Beihang University
yuesenyang@163.com

Naixing Wei
Beihang University
nxwei@buaa.edu.cn

Patterns and meanings of verbal hedges in Chinese scholars English research articles: A corpus-based contrastive approach

Various studies (e.g. Hyland, 1994, 1996a, b, c, 1998a, b; Markkanen & Schröder 1997; Varttala 1999) show that academic discourse is characterized by a heavy use of hedging expressions which serve important roles in helping express a wide variety of evaluative meanings and discourse functions and an appropriate use of this device forms an indispensable part of successful academic communication.

However, most previous studies of hedging expressions and meanings tend to be focused on single words, with insufficient attention paid to the frequent patterns in which hedges occur and the characteristic meanings and functions they express in their local environments. Moreover, while a number of studies deal with the hedging expressions in academic texts written by native scholars (e.g. Hyland 1996b, c, 1998a, b; Myers 1989; Salager-Meyer 1994), little is done about the hedging features of Chinese scholars English research articles (with the exception of Xu, Zheng & Zhang 2014; Yang 2013). To further bridge this gap, this study conducts a corpus-based research of patterns and meanings of verbal hedges in academic texts and aims to explore the following research questions:

(1) What are the frequently occurring patterns of verbal hedges in Chinese scholars and native scholars English research articles? Are there any differences between them in terms of form and frequency?
(2) What specific discourse and strategic meanings are realized by particular lexical sequences of verbal hedges in Chinese scholars English research articles?
(3) What potential implications can be provided for English academic writing for Chinese students and academics?

The corpora analyzed are the Chinese Scholars Academic English Corpus (CSAEC) and the Native Scholars Academic English Corpus (NSAEC), each of which consists of 7.4 million tokens by a random selection of English research articles from international journals with high SCI (SSCI) impact factors in 23 disciplines. The idiom principle (Sinclair, 1991), extended unit of meaning (Sinclair 1996, 2004), pattern grammar (Hunston & Francis 2000) and evaluation (Thompson & Hunston 2000) are adopted as the analytical instruments. The target research words are the 11 most frequent hedging judgmental verbs (INDICATE, SUGGEST, PROPOSE, PREDICT, ASSUME, SPECULATE, SUSPECT, BELIEVE, IMPLY, ESTIMATE, and CALCULATE) listed in Hyland's (1998b) study. We carry out the research in following steps: hedges extraction, pattern identification and function exploration.

This study shows that frequently occurring lexical sequences of verbal hedges have 8 patterns of use. The pattern headed by inanimate nouns is the most frequently used one by both Chinese and native scholars. However, it indicates that the patterns significantly more frequently used by Chinese scholars are We V (that), AUTHOR V (that), It V (that), and V (that), while I V (that) is significantly less frequently used compared with those used by native scholars. In addition, the other two patterns It BE V-ed (that) and Let us V (that) show no strikingly different frequencies in the two corpora.

Corpus evidence from the CSAEC also reveals that various lexical sequences of verbal hedges constitute an integral part of academic texts, and realize specific discourse and strategic meanings. The recurrent lexical sequences of verbal hedges perform five discourse functions in the academic texts under study, namely, interpreting data, formulating claims, validating model or theory, reasoning and reporting, and also realize four strategic meanings: accurate representation strategy, consensus strategy, claim commitment strategy and engagement strategy. Realizations of these

functions depend on the co-selection between lexis, pattern and discourse structure, among others.

The findings of this study have thrown important insights for understanding the feature and nature of hedging patterns in association with academic evaluation and strategy. They are of potential value for improving academic writing pedagogy.

References

Hunston, S., & Francis, G. (2000). Pattern Grammar: A corpus-driven approach to the lexical grammar of English. Amsterdam/Philadelphia: John Benjamins.

Hyland, K. (1994). Hedging in academic writing and EAP textbooks. English for Specific Purposes, 13(3), 239-256.

Hyland, K. (1996a). Nurturing hedges in the ESP curriculum. System, 24(4), 477-490.

Hyland, K. (1996b). Talking to the academy: Forms of hedging in science research articles. Written Communication, 13(2), 251-281.

Hyland, K. (1996c). Writing without conviction? Hedging in science research articles. Applied Linguistics, 17(4), 433-454.

Hyland, K. (1998a). Boosting, hedging and the negotiation of academic knowledge. Text, 18(3), 349-382.

Hyland, K. (1998b). Hedging in Scientific Research Articles. Amsterdam/Philadelphia: John Benjamins.

Markkanen, R., & Schröder, H. (Eds.) (1997). Hedging and Discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts. Berlin/New York: Walter de Gruyter.

Myers, G. (1989). The pragmatics of politeness in scientific articles. Applied Linguistics, 10(1), 1-35.

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. English for Specific Purposes, 13(2), 149-170.

Sinclair, J. M. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sinclair, J. M. (1996). The search for units of meaning. Textus, IX, 75-106.

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

Thompson, G., & Hunston, S. (2000). Evaluation: an introduction. In S. Hunston & G. Thompson (Eds.), Evaluation in Text: Authorial stance and the construction of discourse. Oxford: Oxford University Press, 1-27.

Varttala, T. (1999). Remarks on the communicative functions of hedging in popular scientific and specialist research articles on medicine. English for Specific Purposes, 18(2), 177-200.

Xu, J., Zheng, L., & Zhang H. M. (2014). A corpus-based contrastive study of hedges in mainland Chinese and native scholars' English scientific research articles: A case study of the articles published in Nanotechnology. Foreign Language Learning Theory and Practice, 2(2), 46-55.

Yang, Y. (2013). Exploring linguistic and cultural variations in the use of hedges in English and Chinese scientific discourse. Journal of Pragmatics, 50(1), 23-36.

• • •

Software demonstrations

Vaclav Brezina
Lancaster University
v.brezina@lancaster.ac.uk

Matt Timperley
Lancaster University
m.timperley@lancaster.ac.uk

#LancsBox: A new-generation corpus analysis tool

In this software demonstration, we introduce #LancsBox, a new software package for the analysis of language data and corpora, which was developed at Lancaster University. Following the recent debate in the field (e.g. McEnery & Hardie 2011; Kilgarriff 2012; Gries 2013; Lijffijt et al. 2014; Brezina & Meyerhoff 2014; Brezina et al. 2015; Gablasova et al. 2017) and responding to the challenges identified in the debate, we have developed a software tool that incorporates a number of existing analytical techniques and adds new innovative methods that enable more efficient and sophisticated exploration of the data. #LancsBox can be used by linguists, language teachers, translators, historians, sociologists, educators and anyone interested in quantitative language analysis. It is free to use for non-commercial purposes and works with any major operating system.

#LancsBox takes plain text or XML file input and processes data automatically adding part-of-speech annotation using Tree Tagger (Schmid 1995). It can be used with any language; some types of analysis which require morphological annotation of the data can be performed with languages supported by Tree Tagger. Currently, parameter files are available for Bulgarian, Catalan, Chinese, Coptic, Czech, Dutch, English, Estonian, Finnish, French, Galician, German, Italian, Latin, Mongolian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish and Swahili. However, the users can supply their own parameter files for any language of their choice.

In particular, #LancsBox:

- Searches, sorts and filters examples of language use.
- Compares frequency of words and phrases in multiple corpora and subcorpora.
- Identifies and visualises meaning associations in language (collocations).
- Computes and visualizes keywords.
- Uses a simple but powerful interface.
- Supports a number of advanced features such as customisable statistical measures.

This software demonstration highlights innovative features of the new tool with the focus on visualization of collocations and keywords. In addition, we also discuss more general principles of statistical data analysis and visual data display that can be used for effective presentation of quantitative findings based on language corpora.

#LancsBox can be downloaded for free from the tool website http://corpora.lancs.ac.uk/lancsbox. Version 3 with new features described in this demonstration will be publicly released in September 2017; version 2 is available in the meantime.

References

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

Brezina, V., & Meyerhoff, M. (2014). Significant or random. A critical review of sociolinguistic generalisations based on large corpora. International Journal of Corpus Linguistics, 19(1), 1-28.

Gablasova, D., Brezina, V., & McEnery, A. M. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. Language Learning. DOI: 10.1111/lang.12226

Gries, S. Th. (2013). Statistics for linguistics with R: a practical introduction. Berlin: Walter de Gruyter.

Gries, S. Th. (2006). Some proposals towards a more rigorous corpus linguistics. Zeitschrift für Anglistik und Amerikanistik, 54(2), 191-202.

Kilgarriff, A. (2012). Getting to know your corpus. In Sojka, P., Horák, A., Kopecek, I. & Pala, K. Text, Speech and Dialogue (pp. 3-15). Berlin: Springer.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2014). Significance testing of word frequencies in corpora. Digital Scholarship in the Humanities, advanced access.

McEnery, T., & Hardie, A. (2011). Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

Schmid, H. (1995). Treetagger, a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

• • •

Katharina Ehret
University of Freiburg
katharina.ehret@gmail.com

Marten Juskan
University of Freiburg
marten.juskan@anglistik.uni-freiburg.de

Katja Roller
University of Freiburg
roller.katja@googlemail.com

Bernd Kortmann
University of Freiburg
bernd.kortmann@anglistik.uni-freiburg.de

Exploring English dialects online: The FRED corpus as interactive open-access tool

FRED - the Freiburg Corpus of English Dialects - is a monolingual spoken-language corpus of regional English dialects from nine major dialect areas in England, Wales, Scotland, the Hebrides and the Isle of Man. The corpus samples approximately 2.5 million words of transcribed text and 300 hours of recorded speech (Hernandez 2006; Szmrecsanyi and Hernandez 2007). The central aim in creating FRED was to provide a solid geographically balanced database for investigations into morphosyntactic variation in traditional British English dialects. Until recently, such investigations using the full corpus have been restricted to on-site research in Freiburg, with off-site uses being limited to a small subset of the corpus.

This talk outlines how FRED is now being made available world-wide by publishing it online via the University Library's Current Research Information System (FreiDok plus). We demonstrate how FRED transcripts and audio files (in a first step, of the subset "FRED-S", comprising roughly 1 million words) can be accessed online. Furthermore, we present an interactive research database featuring a full-text search engine and various filters to sort the corpus files by geographical and social parameters, such as dialect area, speaker age and sex. Both the corpus and the database will be linked to a multimedia platform for corpus-based instruction and e-learning ("FREDDIE") which will provide FRED materials for the classroom.

In sum, this data presentation aims to show that FRED online offers multiple opportunities for research, teaching and learning.

References

Hernandez, N. (2006). User's Guide to FRED : Freiburg Corpus of English Dialects. URL http://www.freidok.uni-freiburg.de/volltexte/2489/.

Szmrecsanyi, B. & Hernandez, N. (2007). Freiburg Corpus of English Dialects Sampler („FRED-S"). URL https://www.freidok.uni-freiburg.de/data/2859.

• • •

Gero Kunter
Heinrich-Heine-Universität Düsseldorf
gero.kunter@uni-duesseldorf.de

Coquery: a free corpus query tool

Coquery (http://www.coquery.org) is a free corpus query tool for Windows, macOS, and Linux. Its aim is to provide an accessible working environment for various analysis tasks encountered in corpus linguistics,and for which the available software tools do not offer an easy solution. This software demonstration will give an overview of the main features of Coquery.

In McEnery & Hardie (2012)'s taxonomy, Coquery may be considered a "third-generation concordancer": it is a desktop program which can be used to create concordances, frequency lists, collocations, and keyword analyses (McEnery & Hardie 2012: 41). Yet, it also incorporates many of the characteristics that they ascribe to the web-based "fourth-generation" systems such as a powerful underlying SQL database engine and an optional client-server model (McEnery & Hardie 2012: 45). These characteristics enable the software to manage large corpora such as COCA on a reasonably modern computer.

Coquery supports several English corpora (e.g. BROWN, BNC, COCA, ICE_NG, Switchboard) and lexical data bases (e.g. CELEX, CMUdict). Users can also use Coquery to build their own corpora from text collections in a variety of formats (e.g. HTML, PDF, or .docx). Meta data can be added to each included file, and user corpora can automatically be lemmatized and POS-tagged using the NLTK framework (Bird et al. 2014).

One of the key strengths of Coquery is its user-friendly yet flexible interface which is used to access the supported corpora. The syntax is an extension of the well-known „classic" syntax used by the COCA website. Users can freely choose the corpus features which will be included in the query results. The results can be grouped, sorted (including reverse-sort), and filtered, and users can seamlessly switch between e.g. collocation lists and KWIC views. Different types of functions can be applied to the data, such as string functions (e.g. number of characters, regular expression matching), logical functions (e.g. "larger than", "equal"), and statistical functions (e.g. type-token ratio, query entropy, relative frequency). Data tables from different corpora can be joined using a similar approach to that described in Davies (2007) so that users can, for example, enrich the query matches from the BNC with morphological parses

from the CELEX Lexical Database or phonological transcriptions from the CMUdict pronunciation dictionary. The combination of these features enable users to perform complex analyses without the need of learning a programming language.

Coquery also features a visualization module that encourages the interactive exploration of corpus data in the spirit of Siirtola et al. (2011). For example, the distribution of matches across a corpus can be visualized in barcode plots (similar to those provided in AntConc, Anthony 2005) with interactive controls that provide access to the context of each match. Bar charts and heat maps are available for contingency tables, and diachronic developments may be visualized in different types of time-series plots.

Due to its open-source nature, Coquery welcomes additions from other Python programmers. For example, the list of supported corpora may be expanded by programming new corpus installer modules, or further visualization modules might be added.

References

Anthony, L. (2005). AntConc: a learner and classroom friendly, multi-platform corpus analysis toolkit. Proceedings of IWLeL 2004: An interactive workshop on language e-learning. Tokyo: Waseda University, 7–13.

Bird, S., Klein, E., & Loper, E. (2014). Natural language processing with Python. Retrieved March 10, 2016, from http://www.nltk.org/book/.

Davies, M. (2007). Semantically-based queries with a joint BNC/WordNet database. In Roberta Facchinetti (Ed.), Corpus linguistics 25 years on. Amsterdam: Rodopi, 149–167.

McEnery, T. & Hardie, A. (2012). Corpus linguistics. Cambridge: Cambridge University Press.

Siirtola, H., Nevalainen T., Säily T. & Räihä, K.-J. (2011). Visualisation of text corpora: A case study of the PCEEC. In T. Nevalainen & S. M. Fitzmaurice (Eds.), How to Deal with Data: Problems and Approaches to the Investigation of the English Language over Time and Space. Research Unit for Variation, Contacts and Change in English (VARIENG). Retrieved December 15, 2016, from http://www.helsinki.fi/varieng/series/volumes/07/siirtola_et_al/.

• • •

Daniel McDonald
Eberhard Karls Universität Tübingen
daniel.mcdonald@uni-tuebingen.de

tücan: a tool for more sophisticated corpus linguistics

The increasing number of digitised or born-digital datasets means that computer-aided analysis of texts is becoming more and more common. Corpus linguistics provides a well-established repertoire of tools and practices for performing such analyses: techniques such as keywording, n-gramming and concordancing have aided in diverse tasks such as lexicography, discourse analysis and the development of new grammars.

One shortcoming in contemporary corpus linguistics is a lack of engagement with state-of-the-art methods from natural language processing and computational linguistics, which allow accurate grammatical parsing of arbitrary text, and automated searching of parser output. These parsed data structures facilitate rich insights into discourse, meaning, function and semantics in corpora that are not possible using more traditional corpus linguistic tools. Parsed data makes it possible to move beyond analysis of word-forms and their co-adjacency, toward more sophisticated account of lexicogrammar. Other key NLP tasks, such as co-reference resolution and sentiment analysis, also open up new epistemologies for corpus linguistics. Despite the promise of these developments, tools are still needed in order to make these technologies available to users without a great deal of training in computational workflows.

A second shortcoming is a lack of engagement with the metadata that often accompanies, or is manually added to, digitised text. Corpora can be internally structured according to a metadata feature (genre, date, speaker ID, etc.), so that corpus interrogation can model differences between subcorpora. Dynamic, symbolic arrangement of subcorpora means that the same dataset can be investigated from a number of perspectives. A corpus of news articles containing author names, dates of publication and topics, for example, could be used to model author variation, register or longitudinal language change. At the same time, currently out of reach of current tools is the ability to create new annotations and metadata during the interrogation process. Ideally, it should be possible to locate text spans matching some criteria, annotate these spans with useful information, and then use these annotations as subcorpora or data filters in searches that follow. In this way, it could be possible to collapse the divide between corpus building and corpus interrogation, facilitating iterative improvement of data quality and search accuracy.

In this presentation, I introduce tücan (https://tucan.readthedocs.io), a free, open source, well-documented corpus linguistic tool that addresses the shortcomings listed above. The tool has a powerful backend and a feature-rich, web-based frontend. Each differs in power and ease of use, but both make possible new kinds of functionally driven analysis of text corpora. Both interfaces are presented alongside use-case examples from a project aiming to understanding the shifting use of „risk language“ in print journalism. A roadmap for future development is also briefly discussed.

Harri Siirtola
University of Tampere
Harri.Siirtola@staff.uta.fi

Tanja Säily
University of Helsinki
tanja.saily@helsinki.fi

Terttu Nevalainen
University of Helsinki
terttu.nevalainen@helsinki.fi

Text Variation Explorer 2: a new tool for exploring corpora

Text visualization is an active research area of information visualization. Kucher and Kerren (2015) made a survey of currently published text visualization techniques and maintain an interactive, online directory of them which can be updated by the research community. The number of techniques described has rapidly risen from the initial 141 to 365 in less than two years. Currently, only 7 out of 365 techniques are listed as linguistically motivated, including the first version of our freely available text visualization tool, Text Variation Explorer (TVE; Siirtola et al. 2014). We are developing a new iteration of our tool where the focus is on user-friendly distribution and sampling of corpora.

Interactive text visualization is a way to gain rapid insight into text, its structure, complexity, and variation. This exploration is often a fruitful approach to raise questions and make hypotheses about the text. It is not meant as a substitute for rigorous linguistic or statistical analysis, but to serve as a starting point for a study. Besides text analysis, the interactive approach can also help when we drill down to the part of the corpus we are currently interested in. This phase is usually based on the metadata describing the corpus contents, such as genre, time period or social category. In our tool, a direct-manipulation query interface allows the user to rapidly construct the text samples needed for the problem at hand.

Studying a corpus with an analysis tool often involves many preliminary steps. The corpus text and metadata might need transformations into a format accepted by the tool, like change of character encoding or changes in the POS tagging – steps that many of us find cumbersome. We propose an approach where the corpus is packaged within the analysis tool, ready to be studied. TVE2 is designed to complement concordance-based corpus query software like CQPweb (Hardie, 2013) as an exploratory tool that uses simple, language-independent text measures in an interactive visualization. This approach is especially viable for older, historical corpora, such as the **ICAME Corpus Collection**. These corpora are typically small but carefully compiled. As they have spawned a great deal of research in the past, we wish to introduce them to a new generation of corpus linguists in a format that we believe they will find useful. We also expect the format to prove useful for teaching purposes.

In our software demonstration, we will present TVE2 bundled up with a selection of ICAME corpora, giving them a new life within the software. We will show how TVE2 can be used to discover interesting variation in corpora in terms of several measures, including type/token ratio, proportion of hapax legomena, average word length, and principal component analysis with a user-given list of words. The measures, which are well suited to exploring e.g. stylistic and register variation, are visualized using simple scatterplots, the key feature of which is interaction: users can select the linguistic and metadata variables to be visualized and change them on the fly. Moreover, the visualizations and corpus texts are linked, so that clicking on one will highlight the corresponding section on the others. The new version of our tool makes it easy to compare different time periods, genres and other metadata categories both within and across corpora.

References

Hardie, A. (2013). CQPweb [computer software]. Available from http://cwb.sourceforge.net/cqpweb.php

Kucher, K., & Kerren, A. (2015). Text visualization techniques: Taxonomy, visual survey, and community insights. In S. Liu, G. Scheuermann & S. Takahashi (Eds.), 2015 IEEE Pacific Visualization Symposium (PacificVis). Piscataway, NJ: IEEE, 117-121.

Siirtola, H., Säily, T., Nevalainen, T., & Räihä, K.-J. (2014). Text Variation Explorer: Towards interactive visualization tools for corpus linguistics. International Journal of Corpus Linguistics, 19 (3), 417-429.

• • •

Peter Uhrig
FAU Erlangen-Nürnberg
peter.uhrig@fau.de

A corpus infrastructure for accessing multimodal data: NewsScape and the Distributed Little Red Hen Lab

The UCLA Library Broadcast NewsScape (see Steen & Turner 2013) is a collection of over 300,000 hours' worth of news programmes (in a wide sense, including talk shows) in more than 20 languages. The English portion – mostly from the United States – is by far the largest component with over 330,000 shows totaling at roughly 1.7 billion words in 2016. The collection consists of video recordings aligned with transcriptions extracted from the closed captions (subtitles) transmitted with the video. The transcriptions (all transmitted in upper case) are then automatically annotated with both custom-made tools, e.g. to detect commercials, story boundaries, sentence

boundaries, meta-information such as voice quality, etc. and a range of state-of-the-art NLP tools to obtain PoS, lemma, parse trees, "TrueCased" text, co-reference resolution, named entity recognition, sentiment analysis, frame and semantic role labelling. At the same time, optical character recognition (OCR) is run on the video to capture on-screen text. In a further step, automatic annotation of the video is run to detect elements of the visual representation; a demo of a prototype is available at https://matthewwhitaker.me/gci/realtime/.

The Distributed Little Red Hen Lab (http://www.redhenlab.org) is a global co-operative of researchers working mainly on multimodal communication. In the process, we also contribute to the improvement of the NewsScape collection, e.g. by providing automatic and manual annotation or by deriving corpora to be used in linguistic research. While the co-operative cannot act as a simple service provider it welcomes scholars to join and contribute their annotations, analyses or software in return for access to the collection.

This software demonstration will give an overview of the structure of the NewsScape project and the tools provided the Distributed Little Red Hen Lab, focusing on the English data. The features to be demoed are:

- Corpus queries using
  - a corpus-linguistic approach via a modified version of CQPweb (Hardie 2012) with direct access to videos,
  - a tailor-made search engine that exposes all levels of annotation via a very powerful query builder,
  - a GisGraphy Geo-Information system,
  - the treebank.info interface (Uhrig/Proisl 2012) to search for grammatical structures and collexemes.
- A variety of potential annotation processes for various user requirements
  - within CQPweb,
  - with the Red Hen Rapid Annotator (mainly for classification tasks),
  - with the Online Annotation tool,
  - with standard spreadsheet software, making use of the export/import functionality,
  - with external tools such as ELAN, making use of the export/import functionality.
- Post-query analyses, optionally based on one's own annotations, such as
  - calculation of collocation candidates in CQPweb,
  - analysis of distribution in CQPweb,
  - export to standard spreadsheet software or statistical packages such as SPSS or R.
- Re-integration of one's own annotations into the repository in the spirit of the Distributed Little Red Hen Lab.

References

Hardie, A. (2012) CQPweb – combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics 17 (3), 380-409.

Steen, F. & Turner, M. (2013). Multimodal construction grammar. In M. Borkent, B. Dancygier & J. Hinnell (Eds.), Language and the creative mind. Stanford, CA: CSLI Publications, 255-274.

Uhrig, P. & Proisl, T. (2012). Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates. Lexicographica 28, 141–180.

• • •

Christoph Wolk
University of Giessen
Christoph.B.Wolk@anglistik.uni-giessen.de

Bridgit Fastrich
University of Giessen
Bridgit.C.Fastrich@anglistik.uni-giessen.de

Creating custom concordancers for research and teaching

Most corpus-linguistic work today relies on two major categories of concordancers: corpus-specific concordancers that are tailored to a particular corpus (or set of corpora) and typically hosted on the web, such as the BYU corpus interfaces (e.g. Davies 2008-), and general-purpose concordancers such as AntConc (Anthony 2014) or Wordsmith Tools (Scott 2016), which are designed to work with any corpus and are invaluable because of this flexibility. However, both have significant downsides in areas of research such as corpus-assisted discourse studies (see Baker 2006), for which a specialized corpus must often be constructed. This is because preexisting corpus-specific concordancers typically cannot be used with other corpora, and general-purpose concordancers are somewhat limited by their generality. They are designed to work best for longer prose texts and often cannot leverage the structure of different kinds of corpus data, such as paired elements in a question-and-answer corpus or lines of dialogue in a TV script corpus. Furthermore, specialized corpora often have rich, multi-level annotations, such as topic or question type. Researchers may find it helpful to be able to quickly and interactively restrict particular corpus-linguistic analyses to only certain parts of the data, but this is a somewhat burdensome process in general concordancers. Another problematic aspect of general-purpose concordancers is that they may be confusing to learn for beginning users. This can make it difficult to inte-

grate corpus-based practices into general linguistics and ESP teaching, which would otherwise be a natural fit for corpus methodologies.

We argue that corpus-specific custom concordancers can alleviate the above concerns. First, they can be tailored to the structure of the (meta)data and therefore enable convenient access to the data for research purposes. Second, they work well in the classroom: much of the incidental complexity of corpus work, such as loading corpora, is already taken care of, and instructors can adapt the capabilities of the tool to the needs and goals of the class, leading to a straightforward and clean interface.

To this end, we introduce an open-source, cross-platform (Windows, Mac, Linux, mobile) framework for building custom concordancers. This framework is built with the package Shiny (Chang et al. 2016) for the programming language R (R Core Team 2016) and yields concordancers that can be hosted on the web or run on local computers. R is frequently used as a statistics toolkit within linguistics departments, so the construction of custom concordancers can leverage existing expertise. Our primary aim, however, is to allow even users with minimal programming knowledge to construct simple concordancers for specific kinds of data . We show how to use the concordancers and framework, and showcase some example applications.

References

Anthony, L. (2014). AntConc (Version 3.4.3). Tokyo, Japan: Waseda University.

Baker, P. (2006). Using corpora in discourse analysis. London: Continuum.

Chang, W., Cheng, J., Allaire, J.J., Xie, Y. & McPherson, J. (2016). shiny: Web Application Framework for R. R package version 0.14.2.

Davies, M. (2008-). The Corpus of Contemporary American English: 520 million words, 1990-present.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Scott, M. (2016). WordSmith Tools (Version 7). Stroud: Lexical Analysis Software.

• • •