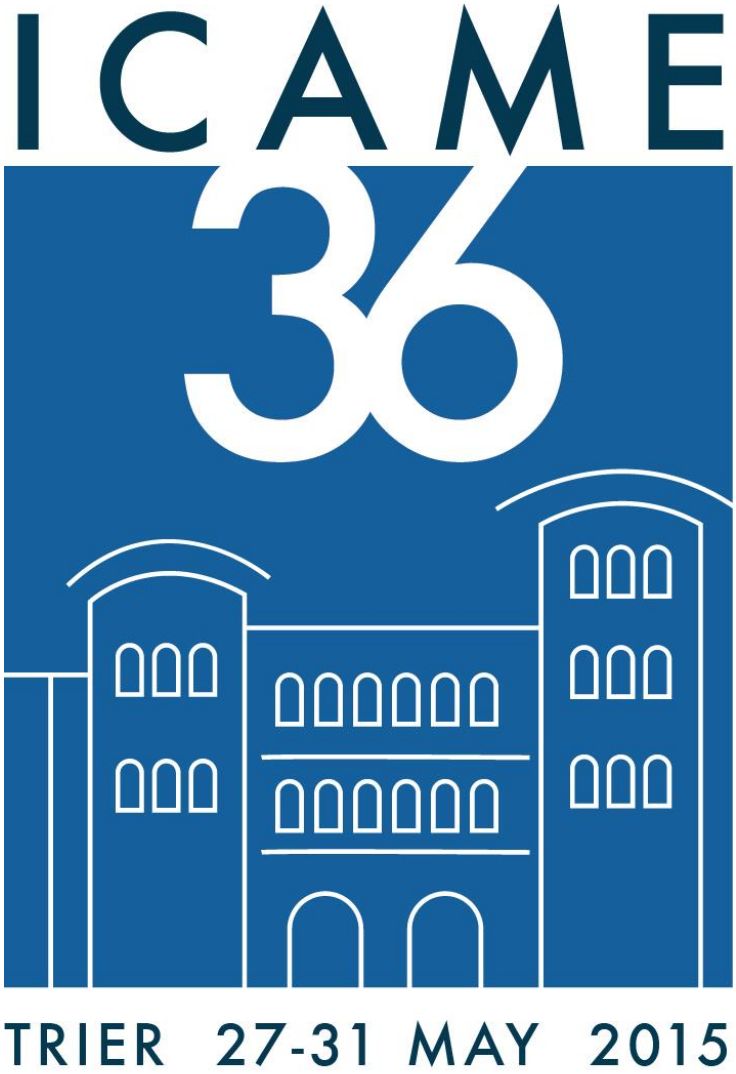


WORDS, WORDS, WORDS – CORPORA AND LEXIS

ABSTRACTS

Table of Contents

Welcome to ICAME 2015 in Trier!

We sincerely hope that you will all enjoy every aspect of this conference. Please let one of our team of organisers and helpers know if there is anything we can do to improve your stay.

We hope you will enjoy exciting plenaries, top-class research papers, excellent wine, delightful companionship and conversation.

Organizing Committee:

- Sabine Arndt-Lappe
- Anne-Katrin Blass
- Daniela Kolbe-Hanna
- Sebastian Hoffmann (chair)
- Lilian Lee Hoffmann
- Kerstin Lunkenheimer
- Andrea Sand
- Michael Stubbs

Student Helpers:

- Lisa Dillmann
- Denis Gusakov
- Franziska Hackhausen
- Jenifer Ramberger
- Isabelle Reinhardt

Pre-conference workshops

Workshop I: Are there monolingual corpora? (Computer says no.)

Convenors: Arja Nurmi & Päivi Pahta (University of Tampere)

Beyond the monolingual ideal? Advanced L2 English(es) on a diachronic continuum

Mikko Laitinen (Linnaeus University)

This presentation discusses new corpus-based ways of approaching advanced L2 English(es) in present-day multilingual settings. The broad context is the diversification of English today (Schneider 2014) when it is used as an additional linguistic resource alongside various L1s. This expansion has already led some scholars (Mukherjee & Hundt, eds. 2011) to reconsider the partly monolingual ideal in English corpus linguistics, and the presentation aims at contributing to the debate. The monolingual ideal is understood as (a) dividing the sources of evidence into neat sets of native, second language and foreign language corpora (the corpus level), (b) drawing clear boundaries between what actually constitutes a language (e.g. the lexical level), (c) focusing on norms and models of usage (e.g. the lexico-grammatical level).

Parallels are first made to other fields in which the monolingual ideal has already been questioned, either directly or directly. Cases in point are for instance the study of English as a lingua franca (ELF) where the focus falls not on norms of use, but on creative practices in which meanings are actively negotiated by non-native speakers (Seidlhofer 2011; Jenkins et al. 2011; Mauranen 2012) or recent advances in ethnographically-oriented research on multilingualism (Jørgensen et al. 2011; Blommaert & Rampton 2011).

The presentation then introduces ongoing research on advanced L2 English. Its aim is to test the applicability of some of the methods from diachronic linguistics in understanding the globalization of English and to chart new ways of approaching lexico-grammatical variability in it. This research sees advanced L2 English as one stage in the long continuum of Englishes and investigates how English is shaped by speakers/writers in today's multilingual settings (cf. Mair 2013) and how they adapt to variability in the core grammar (cf. Laitinen & Levin forthcoming). The approach suggested is new since previous approaches both in learner English research and ELF have been synchronic. It should be noted, however, that a related approach has recently been elaborated in the study of post-colonial World Englishes by Noël, Van Rooy & van der Auwera (2014).

The presentation zooms into ongoing corpus compilation work in which the aim is to collect a representative sample of English texts in multilingual settings. The objective is that the sampling frame should enable diachronic and diatopic analyses of advanced L2 use and make possible quantitative comparisons between advanced L2 evidence and some of the existing standard English corpora. Finally, I will introduce the set of research topics, stemming from

the corpus material, which have so far been investigated, and discuss a set of broader research questions on which this type of corpus material of English texts in multilingual settings could shed more light.

References

Blommaert, J. & B. Rampton. 2011. Language and superdiversity. *Language and Superdiversities,* 13:2: 1-22.

Jenkins, J., A. Cogo & M. Dewey. 2011. Review of developments in research into English as a lingua franca. *Language Teaching,* 44(3): 281-315.

Jørgensen, J.-N., M. Karrebæk, L. Madsen & J. Møller. 2011. Polylanguaging in superdiversity. *Language and Superdiversities,* 13(2): 23-37.

Laitinen, M. & M. Levin. Forthcoming. On the globalization of English: observations of subjective progressives in present-day Englishes. In E. Seoane & C. Suárez-Gómez eds. *World Englishes: New Theoretical and Methodological Considerations* (*Varieties of English around the World*). Amsterdam: Benjamins.

Mair, C. 2013. The World System of Englishes. Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide*, 34:3: 253-78.

Mauranen, A. 2012. *Exploring ELF: Academic English Shaped by Non-Native Speakers*. Cambridge: Cambridge University Press.

Mukherjee, J. & M. Hundt eds. 2011. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: Benjamins.

Noël, D., B. van Rooy & J. van der Auwera. 2014. Diachronic approaches to modality in World Englishes: introduction to the special issue. *Journal of English Linguistics*, 42(1): 3-6.

Schneider, E. W. 2014. New reflections on the evolutionary dynamics of world Englishes. *World Englishes,* 33(1): 9-32.

Seidlhofer, B. 2011. *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

…sɛdeɛ ɛbɛyɛ a after the court proceedings no nso Ghana bɛkɔso a enjoy asomdwoeɛ a yɛre enjoy yi – Exploring Akan-English code-switching in a corpus of Akan broadcast discussions.

Gloria Otchere (University of Oslo)

The centuries-old contact between the English language and Akan, the most widely spoken indigenous language in Ghana, has had a significant influence on English in Ghana. English in Ghana is currently ranked at the Nativisation-Endonormative stage on Schneider's model for the development of second language varieties of English; an indication of the localization of the English language in Ghana with very conspicuous traces of the indigenous Ghanaian languages. It is, however, obvious that the influence is bi-directional as contemporary spoken Akan exhibits widespread borrowings from English and code-switches between Akan and English. A typical example is the title of this paper which is an extract from an Akan corpus of broadcast discussions. While some attempts have been made at exploring the influences of Akan on English, very little has been done to explore the influence of English on Akan. This paper is a step in this direction. It explores the morphosyntactic rules governing Akan-English code-switches from an insertion model perspective. Specifically, it aims at identifying both the syntactic frames or matrix and the types of elements that are or may be inserted in instances of Akan-English code-switches.

Previous explorations of Akan-English code-switching have shown, among others, evidence for both inter and intra sentential code-switching (Forson, 1979) and that it is socioeconomically, religiously and politically constrained (Albakry & Ofori, 2011; Nartey, 1982; Nyarko, 2012). None has focused on the grammar of Akan-English code-switching. As such, not much is known about the patterns of embedded English language insertions into Akan frames. The study adopts the insertion model of code-switching (cf. Myers-Scotton, 1993, 1995, 1997) which have become very particularly influential in exploring grammatical regularities in code-switching.

The data for the study is a corpus of Akan radio broadcast discussions. It comprises discussions held between 2013 and 2014 in one of the Ashanti regional branch of the Ghana Broadcasting Corporation. It thus contains mainly the Asante Twi dialect of Akan.

It is expected that the study will shed light on the morphosyntactic regularities of Akan-English code-switching.

References

Albakry, M. A., & D. M. Ofori. 2011. Ghanaian English and code-switching in Catholic churches. *World Englishes*, 30(4), 515-32.

Forson, B. 1979. *Code-Switching in Akan-English Bilingualism.* Unpublished PhD, UCLA.

Myers-Scotton, C. 1993. *Dueling Languages. Grammatical Structure in Code-Switching.* Oxford: Claredon.

Myers-Scotton, C 1994. Language processing and the mental lexicon in bilinguals. In R. Dirven & J. Vanparys eds. *Current Approaches to the Lexicon. A Selection of Papers Presented at the 18th LAUD Symposium, Duisburg, March 1993*. Frankfurt: Peter Lang, 73-100.

English as a lingua franca and multilingual practices: A corpus compiler's perspective

Marie-Luise Pitzl (University of Vienna)

Building a corpus of spoken English as a lingua franca (ELF) in which participants with various L1 backgrounds interact in different contexts poses a range of challenges to corpus compilers. During the process of corpus compilation, these challenges occur at different stages and relate to different aspects, such as corpus structure, transcription, mark-up, annotation and metadata (see e.g. Breiteneder et al. 2006; Breiteneder et al. 2009; Majewski 2011; Osimk-Teasdale 2015; Pitzl et al. 2008; Seidlhofer 2001). All of these (and the decisions taken by compilers) influence the usability of the corpus afterwards, so methodological considerations have a long-lasting effect on the kind of research that can be conducted with the corpus in the decades to follow. One of the challenges that comes up repeatedly when building an ELF corpus is how – and to what extent – to represent speakers' use of languages other than 'English'. Code-switching and multilingual practices are an integral part of English as a lingua franca (e.g. Cogo 2012; Klimpfinger 2009); there is a hardly an ELF speech event that is 'monolingual English only'. So any ELF corpus has to come to terms with the multilingual nature of the data methodologically and practically.

This presentation provides an inside-view on how questions concerning the representation of multilingual practices of ELF speakers were resolved in the *Vienna-Oxford International Corpus of English* (VOICE), i.e. the first general ELF corpus to be made publicly available in 2009. Tracing the steps of the compilation process, the presentation reports on decisions taken in light of the multilingual nature of ELF during the transcription of audio material. It introduces the subsequent encoding of 'non-English speech' in TEI-based XML mark-up, paying attention not only to the main body of transcripts but also to metadata and external points of reference. It will also briefly comment on annotation decisions that were taken concerning so-called 'foreign words' during the part-of-speech (POS) tagging of VOICE. For each of these areas, the presentation will provide examples and illustrate the main reasons for the encoding strategies that were adopted. Drawing on a case study of functions of multilingual practices (e.g. inclusion and exclusion strategies) in an ELF business meeting, the presentation will exemplify the potential – but also point out limitations – that VOICE has for the analysis of code-switching and multilingual practices in ELF interactions.

References

Breiteneder, A., T. Klimpfinger, S. Majewski & M.-L. Pitzl. 2009. The Vienna-Oxford International Corpus of English (VOICE). A linguistic resource for exploring English as a lingua franca. *ÖGAI Journal*, 28(1): 21-26.

Breiteneder, A., M.-L. Pitzl., S. Majewski & T. Klimpfinger. 2006. VOICE recording - methodological challenges in the compilation of a corpus of spoken ELF. *Nordic Journal of English Studies*, 5(2): 161-88.

Cogo, A. 2012. ELF and super-diversity: a case study of ELF multilingual practices from a business context. *Journal of English as a Lingua Franca*, 1(2): 287-313.

Klimpfinger, T. 2009. 'She's mixing the two languages together' – Forms and functions of code-switching in English as a lingua franca. In A. Mauranen & E. Ranta eds. *English as a Lingua Franca: Studies and Finding.* Newcastle upon Tyne: Cambridge Scholars Publishing, 348-71.

Majewski, S. 2011. *Design and Implementation of a Research Infrastructure for a Corpus of Spoken ELF*. Vienna: University of Vienna MA.

Osimk-Teasdale, R. 2015. *Parts of Speech in English as a Lingua Franca: the POS Tagging of VOICE*. Vienna: University of Vienna PhD.

Pitzl, M. L., A. Breiteneder & T. Klimpfinger. 2008. A world of words: processes of lexical innovation in VOICE. *Vienna English Working Papers*, 17(2): 21-46.

Seidlhofer, B. 2001. Closing a conceptual gap: the case for a description of English as a lingua franca. *International Journal of Applied Linguistics,* 11(2): 133-58.

Wetin dey happen? Digging into the progressive aspect in multilingual settings

Paula Rautionaho (University of Tampere)

This paper sets out to examine progressive constructions in settings where English co-exists with one or more languages, focusing on constructions other than the standard BE + V*ing* (as in *I'm reading Treasure Island*). Earlier research on a number of varieties represented by components of *the International Corpus of English* (e.g. Rautionaho 2014) has shown that the number of occurrences of BE + V*ing* in World Englishes (such as BrE, IrE, IndE, HKE)

varies a great deal, in fact, to the extent that the number of progressives is almost tripled when considering the varieties with the lowest and the highest frequencies.

This fact leads us to the important question of whether every possible progressive token was indeed captured during the retrieval process. In fact, evidence suggesting that the search strings used may have been incomplete is encountered in Rautionaho (2014) regarding Hong Kong English: the data includes a number of instances that could be analysed as indicating progressive aspect although the form differs from the standard construction (possibly due to influence from Cantonese). Similarly, it seems that the multilingual settings of countries such as Nigeria and Jamaica have led to a situation in which there are other ways of expressing the progressive aspect. In Hong Kong, Nigeria and Jamaica, English is used alongside native languages and/or on a continuum with a creole (Cantonese, Nigerian Pidgin English and Jamaican Creole, respectively).

The present paper thus investigates other possibilities of expressing the progressive aspect in HKE (e.g. BE + V), NigE (e.g. *dey* + V) and JamE (e.g. *a* + V). While doing so, special attention is paid to the design of the spoken part of the ICE corpora as regards the presence of languages other than English, or of varieties of English other than the one in focus. Are the ICE corpora monolingual, or rather, mono-varietal? What are the implications of the compilers' attempt to depict the *standard* forms of World Englishes (Greenbaum 1988)? The fluctuation of the frequency of the progressive tokens in different ICE components will also be addressed: ICE-Nig does indeed contain a number of non-standard progressive constructions, while the same cannot be said of ICE-Jam. For this reason, the *Corpus of Web-based Global English* (Davies 2013) will also be consulted in order to investigate the alternate forms of expressing the progressive aspect.

References

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

Greenbaum, S. 1988. A proposal for an international computerized corpus of English. *World Englishes* 7(3): 315.

Rautionaho, P. 2014. *Variation in the Progressive: A Corpus-Based Study into World Englishes.* Tampere University Press: Acta Universitatis Tamperensis: 1997

Approaching the multilingual realities in the *Corpus of English Religious Prose*

Tanja Rütten (University of Cologne)

The *Corpus of English Religious Prose* (COERP) makes accessible one of the most fundamental discourse domains in the recorded linguistic history of English. Sermons, prayers, religious tracts, and devotional writings in the shape of hagiographies and martyr stories present one of the largest, widely read and most robust body of writings that informs us about the linguistic history of roughly 1,500 years of English. Or does it?

For the greater part of this period, English was only one of a number of languages composers, compilers and copyists of religious texts in England knew and used. While it is a truism that the clergy knew Latin, and sometimes also Greek and/or Hebrew, we tend to ignore this multilingual reality of the authors whose texts we use as data when we investigate earlier *English*.

In compiling COERP, we aimed at representing the variety of codes that occur in the material in various ways. In my contribution, I focus on two issues connected to multilingualism. The first relates to the selection and presentation format of multilingual material. I discuss the question whether it is sensible to confront linguists who are interested in the English language with Latin, Greek, Hebrew, and the occasional Arabic in the first place. If so, can these codes be presented in a linear fashion within the running English text? Should they? Or is it more appropriate to assign them the status of "sub-text" that occurs by request of the researcher in the digital file? And if we see them as an additional layer of the English text at best, how does this affect the monolingual, "raw" English text and our perception of it? Are we even aware of possible intertext from the Bible and other sources?

The second issue I will address relates more immediately to the multilingual influence of our "informants". In a pilot survey of the English subjunctive, I show how its presence in Middle and Early Modern English texts in COERP is often enforced by the Latin sub-text, i.e. the bilingual or multilingual status of the authors. This influence vitally demonstrates the relevance of representing languages other than English in (historical) corpora, since the "bilingual sub-text" in this case is crucial for our understanding of linguistic change in English. I will show how in COERP such sub-text can be mapped onto the respective English passages, in order to retrieve this multilingual influence directly and systematically.

Multilingual elements and mise-en-page in early modern sermons from the *Lampeter Corpus*

Jukka Tuominen (University of Tampere)

Sermons, like other religious texts, had tremendous socio-cultural importance in early modern England, and their relatively stable communicative setting makes them a viable object for diachronic study (Kohnen 2010). They have also been identified as one of the most typical genres in early English to show code-switching, the use of two or more languages in a single communicative event (e.g. Schendl 1996; Pahta and Nurmi 2006). Situated in the middle ground of the continuum from speech to writing, sermons have been characterized e.g. as "scripted" (Pahta and Nurmi 2006) or "speech-purposed" (Culpeper and Kytö 2010), with an implied emphasis on the oral delivery. Since research on early modern pulpit oratory tends to rely on printed source material, analysis of the data requires a careful assessment of the various stages of text production, as well as an awareness of genre and publishing conventions. How were the needs and expectations of the two audiences – the congregation who first heard the sermon and the potential wider readership of the published text – taken

into account (cf. Groeger 2010)? Whose linguistic choices does the final printed version of a sermon ultimately represent?

Recent years have seen a growing scholarly interest in the "visual pragmatics" (Machan 2011) of multilingualism in both past and present-day writing. Foreign words and longer passages appear in ostensibly English texts both within the body and as part of various paratextual elements, such as dedications and marginalia (cf. Genette 1997). How might these aspects of multilingual texts be studied from a corpus-linguistic perspective? Coding that takes into account the variety of languages in a text as well as its macro-level structure provides means for analyzing multilingual discourse practices both quantitatively and qualitatively. Using ten sermons published in pamphlet form between 1640 and 1740 from the *Lampeter Corpus of Early Modern English Tracts* (Schmied et al. 1998), the paper examines the distribution of multilingual elements on the page and across the text as a whole. Differences in the functional types, glossing, and layout-related features of code-switching into Latin, Greek and Hebrew in the material bring out the communicative ends of early modern pulpit oratory and its associated literacy practices, including the layers of production reflected in the printed text.

References

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

Genette, G. 1997. *Paratexts: Thresholds of Interpretation*. Trans. J. E. Lewin. Cambridge: Cambridge University Press. (Originally published in French as *Seuils*, 1987.)

Groeger, D. 2010. *The Pamphlet as a Form of Publication: A Corpus-Based Study of Early Modern Religious Pamphlets*. Aachen: Shaker Verlag.

Kohnen, T. 2010. Religious discourse. In A. H. Jucker & I. Taavitsainen eds. *Historical Pragmatics*. Berlin & New York: De Gruyter Mouton. 523-47.

Machan, T. W. 2011. The visual pragmatics of code-switching in late Middle English literature. In H. Schendl & L. Wright eds. *Code-Switching in Early English*. Berlin & Boston: De Gruyter Mouton. 303-33.

Pahta, P. & A. Nurmi .2006. Code-switching in the *Helsinki Corpus*: a thousand years of multilingual practices. In N. Ritt et al eds. *Medieval English and its Heritage*. Frankfurt am Main: Peter Lang; 203-20.

Schendl, H. .1996. Text types and code-switching in medieval and Early Modern English. *VIEWS* 5 (1–2): 50–62.

Schmied, J., C. Claridge and R. Siemund (compilers). 1998. *The Lampeter Corpus of Early Modern English Tracts*. Chemnitz: [Chemnitz University of Technology]

Workshop II: Corpus linguistics and linguistic innovations in non-native Englishes

Convenors: Sandra C. Deshors (Las Cruces), Sandra Götz (Giessen) & Samantha Laporte (Louvain-la-Neuve)

"It's always different when you look something from the inside": Linguistic innovation in a corpus of EFL Skype conversations

Marie-Louise Brunner, Stefan Diemer & Selina Schmidt (Saarland University)

The paper will discuss linguistic creativity in informal academic Skype conversations between non-native speakers of English. The basis for the study is the Corpus of Academic Spoken English (CASE, cf. Diemer et al. forthcoming), a corpus of more than 200 hours of Skype conversations between L2 speakers of English, currently being compiled at Saarland University, Germany, with partners from Bulgaria, Spain, Italy, Finland, and an L1 reference component from the UK. The paper examines to what extent non-standard features can be identified as examples of language innovation in an English as a Foreign Language (EFL) context in CASE. We will analyze whether participants' language use goes beyond the basic need to make themselves understood and points towards a more assertive and creative perspective on language, reflecting (inter)cultural influences. The paper will also examine whether participants accommodate to each others' innovative language use (cf. Mauranen 2012), thus creating their own ephemeral variety.

The study uses n-gram analysis and word frequency counts in combination with qualitative analysis to extract non-standard features from the corpus data. Problems encountered during the collection process and difficulties in identifying creative use will be discussed. In contrast to Croft's (2000) and van Rooy's (2011) distinction between "error" and "conventionalized innovation" based on diffusion as a criterion, we argue for a more flexible approach, following Kachru's (2006: 247 f.) distinction between "errors" and "functionally appropriate innovation." We thus propose to conceive of non-standard forms as being non-innovative morphosyntactical deviations or functionally accepted lexical innovations, such as new word formations, hybridization, collocations and idioms.

Non-standard language use by EFL speakers in CASE seems to show similar patterns of innovation to English as a Second Language (ESL) varieties (cf. Crystal 2006, Kachru 2006), as has previously been discussed in corpus-based contrastive case studies by Gilquin (2011), Laporte (2012), and Deshors (2014), among others. The strategies used in the EFL setting do not seem to create a separate variety (or separate varieties), although they may be similar and potentially influenced by the interacting native languages.

A first analysis seems to indicate that EFL speakers in an international context often do not thematize non-standard modifications or innovations (for example using other-(initiated)

repair, cf. Levinson 1983). Creative language use, and code-switching, might even have a positive effect on the communicative setting and promote rapport between interlocutors (Spencer-Oatey 2000).

The use of these strategies also emphasizes participants' own cultural identity (Auer 2005, Auer & Eastman 2010). The study shows that a corpus-based analysis is well suited for isolating and illustrating linguistic innovation in an EFL context. Similar to the establishment of New Englishes, the observed EFL communicative adaptations should not be considered defective but rather as enriching and enhancing the communicative properties of English in an international context.

References

Auer, P. 2005. A postscript: code-switching and social identity. *Journal of Pragmatics,* 37: 403-10.

Auer, P. & C. M. Eastman. 2010. Code-switching. In J. Jaspers, J.-O. Östman & J. Verschueren eds. *Handbook of Pragmatics,* vol 7: *Society and Language Use*, 84-112. Amsterdam: Benjamins.

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach.* London: Pearson Education.

Crystal, D. *English as a Global Language.* Oxford: Oxford University Press.

Diemer, S., M.-L. Brunner, C. Collet & S. Schmidt. Forthcoming. *CASE: Corpus of Academic Spoken English.* Saarbrücken: Saarland University (coordination) / Sofia: St Kliment Ohridski University / Forlì: University of Bologna-Forlì / Santiago: University of Santiago de Compostela. http://www.uni-saarland.de/index.php?id=36728. Accessed 28 Jan 2014.

Deshors, S. C. 2014. A case for a unified treatment of EFL and ESL: a multifactorial approach. *English World-Wide,* 35(3): 279-307.

Gilquin, G. 2011. Corpus linguistics to bridge the gap between World Englishes and Learner Englishes. C*ommunicación en el siglo XXI* vol. II, Centro de Lingüística aplicada: Santiago de Cuba: 638-42.

Kachru, B. B. 2006. The English language in the outer circle. *World Englishes*. London: Routledge (2006): 241-255.

Laporte, S. 2012. Mind the gap! Bridge between World Englishes and Learner Englishes in the making. *English Text Construction* 5(2): 265-92.

Levinson, S. C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Mauranen, A. 2012. *Exploring ELF: Academic English Shaped by Non-Native Speakers.* Cambridge: Cambridge University Press.

Spencer-Oatey, H. 2000. Culturally speaking: managing rapport through talk across cultures. In H. Spencer-Oatey ed. *Language, Culture and Rapport Management*. London: Continuum. 1-10.

Van Rooy, B. 2011. A principled distinction between error and conventionalized innovation in African Englishes. In M. Hundt & J. Mukherjee eds. *Exploring Second-Language and Learner Englishes: Bridging the Paradigm Gap*. Amsterdam: Benjamins. 189-208.

Towards a process-oriented approach to comparing EFL and ESL varieties: A corpus-study of lexical innovations

Marcus Callies (University of Bremen)

In the context of current research in the field of corpus linguists that challenges the traditional division between foreign language / learner varieties of English (EFL) and institutionalized second-language varieties of English (ESL) (see e.g. Gilquin 2015), this paper presents a

comparative corpus-study of lexical innovations and creative coinages in derivational morphology. Written EFL data from the *International Corpus of Learner English* (ICLE; Granger et al. 2009) will be compared to similar ESL data from the *International Corpus of English* (ICE; Greenbaum 1996). The paper adopts a process-oriented approach to comparing EFL and ESL varieties and examines to what extent they are driven by general cognitive processes of language acquisition such as simplicity, regularity, analogy, or isomorphism (Sharma 2012, Schneider 2012).

The testing ground for examining these two types of varieties is word formation, a major mechanism for the expansion of the vocabulary in a language that involves knowledge of the combinatory properties of affixes and bases. There is surprisingly little research on EFL learners' productive use of derivational morphology. Similarly, studies on word-formation in ESL varieties have been rather sparse (mostly lacking quantitative documentation), and have focused on selected varieties and descriptive accounts of a small number of productive nominal and adjectival suffixes (see e.g. Biermeier 2009, 2014).

The data evidence similar types of innovations in EFL and ESL varieties, e.g. manifestations of overgeneralization of affixation where it does not apply in Standard English because of conversion or subtractive processes ("overaffixation") and other "overexplicit" forms motivated by isomorphism in that additional (or more abstract) meaning is marked by additional linguistic form, i.e. nominal and verbal suffixes. This confirms recent studies that have suggested that overgeneralizations and paradigmatic formations (e.g. back-formations from more complex forms) play an important role in learner varieties, while processes such as conversion or subtraction of form seem to be dispreferred (e.g. Plag 2009, Callies 2015).

More generally, the findings of the present study also confirm several previous observations that cognitively motivated processes functioning to maximize transparency and increase explicitness are at play in ESL and EFL varieties (Schneider 2012: 67f.): the use of *to*-infinitive complements with causative *make* (Laporte 2012) and phrasal/prepositional verbs in which semantically redundant particles are used to explicitly mark the directionality that is implicit in the verb (*enter into, return back, surface up, rise up*; Nesselhauf 2009, Gilquin 2015).

References

Biermeier, T. 2009. Word-formation in New Englishes. Properties and trends. In T. Hoffmann & L. Siebers eds. *World Englishes. Problems, Properties and Prospects*. Selected papers from the 13th IAWE conference. Amsterdam: Benjamins, 331-49.

Biermeier, T. 2014. Compounding and suffixation in World Englishes. In S. Buschfeld, T. Hoffmann, M. Huber & A. Kautzsch eds. *The Evolution of Englishes: The Dynamic Model and Beyond*. Amsterdam: Benjamins, 312-30.

Callies, M. 2015. Effects of cross-linguistic influence in word formation. A comparative learner-corpus study of advanced interlanguage production. In H. Peukert ed. *Transfer Effects in Multilingual Language Development*. Amsterdam: Benjamins, 127-43.

Gilquin, G. 2015. At the interface of contact linguistics and second language acquisition research. New Englishes and Learner Englishes compared. *English World-Wide*, 36(1), 90-123.

Granger, S., E. Dagneaux, F. Meunier, & M. Paquot. 2009. *The International Corpus of Learner English*, Version 2. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.
Greenbaum, S. ed. 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.
Laporte, S. 2012. Mind the gap! Bridge between World Englishes and Learner Englishes in the making. *English Text Construction*, 5(2): 265-92.
Nesselhauf, N. 2009. Co-selection phenomena across New Englishes. *English World-Wide*, 30(1): 1-26.
Plag, I. 2009. Creoles as interlanguages: word-formation. *Journal of Pidgin and Creole Languages*, 24(2): 339-62.
Schneider, E. 2012. Exploring the interface between World Englishes and Second Language Acquisition – and implications for English as a Lingua Franca. *Journal of English as a Lingua Franca*, 1, 57-91.
Sharma, D. 2012. Second language varieties of English. In T. Nevalainen & E. Cloos Traugott eds. *The Oxford Handbook of the History of English*. Oxford: OUP, 582-91.

Conversion in Asian Englishes: Are Singapore English and Hong Kong English really ESL varieties?

Stephanie Horch (University of Freiburg)

Singapore English and Hong Kong English have both been classified as ESL varieties in the Kachruvian Three Circles model (1985). Both varieties show a similar contact ecology (Chinese), but differ in their socio-institutional status in Schneider's Dynamic Model (2007). In a study of conversion in these two varieties, I show that despite the obvious similarities in substratum, the usage frequency of conversion in both varieties differs considerably. These findings, similar to –most recently– Deshors' (2014) and Gilquin's (2015), call into question the established notion of ESL and the status of SgE and HKE as ESL varieties.

Conversion, a process situated at the lexis-grammar interface, has not received much attention yet. This is in part due to the comparatively small size of corpora such as ICE (used in Biermeier 2008). Using the *Corpus of Global Web-based English* (Davies 2013), verb-to-noun conversion in selected Asian Englishes, with a focus on HKE and SgE, is analyzed to investigate how transfer from the substratum and institutionalization interact to yield distinctive usage patterns.

For each variety, normalized ratios of potentially converted nouns (e.g. *a require*) to synonymous, derived nouns (e.g. *a requirement*) are determined on the basis of random samples from GloWbE. A logistic regression reveals that both HKE and SgE show higher odds of conversion than British English, the parent variety. Nonetheless, in HKE, this trend is more pronounced than in SgE. Furthermore, in HKE, the odds of conversion increase for less frequent verbs. That is, in the low-frequency range, speakers fall back on a pattern from their L1, following the shortest path principle (Biewer 2011).

A qualitative analysis of data from GloWbE and ICE (The ICE Project 2014) complements the quantitative findings and shows that conversion is more frequent in HKE in formal

contexts such as legal cross-examinations or business transactions, while it is more likely to be found in personal interaction in SgE. The spoken data further reveal that conversion in SgE is followed by self-repair (Schegloff et al. 1977), hinting at a high language awareness on the part of the speaker. In HKE, on the other hand, conversion in speech seems to go without noticing. This shows that HKE speakers, despite an exonormative orientation towards the BrE standard (Pang 2003), do not show the same awareness as speakers of SgE.

The results suggest that in conversion, effects of the substratum are not everything. The degree of institutionalization in contact dialect formation is crucial, as a high degree of institutionalization reduces both the quantity and range of transfer, regardless of the official status of English in the region. Even though SgE and HKE have both been termed ESL varieties, SgE presents a much higher degree of institutionalization than HKE. Subsuming SgE and HKE under the heading of ESL varieties consequently does not truthfully represent language use. This ultimately necessitates that the notion of ESL rather be understood as (part of) a continuum.

References

Biermeier, T. 2008. *Word-Formation in New Englishes. A Corpus-Based Analysis*. Münster: Lit.

Biewer, C. 2011. Modal auxiliaries in second language varieties of English: A learner's perspective. In J. Mukherjee & M. Hundt eds. *Exploring Second Language Varieties of English and Learner Englishes. Bridging a Paradigm Gap*. Amsterdam: Benjamins, 7-34.

Davies, M. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries. http://corpus2.byu.edu/glowbe/.

Deshors, S. C. 2014. A case for a unified treatment of EFL and ESL. A multifactorial approach. *English World-Wide*, 35: 277-305.

Gilquin, G. 2015. At the interface of contact linguistics and second language acquisition research. New Englishes and Learner Englishes compared. *English World-Wide*, 36: 91-124.

Kachru, B. B. 1985. Standards, codification and sociolinguistic realism: the English language in the Outer Circle. In R. Quirk & H. G. Widdowson eds. *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11-30.

Pang, T. T. T. 2003. Hong Kong English: A stillborn variety? *English Today*, 19: 12-18.

Schegloff, E. A., G. Jefferson & H. Sacks. 1977. The preference for self correction in the organization of repair in conversation. *Language*, 53: 361-82.

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

The ICE Project. 2014. International Corpus of English. http://icecorpora.net/ice/index.htm.

"This hair-style called as 'duck tail'": Tracing the 'intrusive *as*' construction in South Asian varieties of English and learner Englishes

Christopher Koch, Claudia Lange & Sven Leuckert (Dresden University of Technology)

It has been observed (e.g. Nihalani et al. 1979: 276, Yadurajan 2011) that Indian English differs from other varieties of English in that it allows 'intrusive *as*' in complex transitive constructions. Early mentions of the feature date back at least as far as 1934 (Goffin 1934). However, after initial empirical observations on its spread in Sedlatschek (2009), Lange's

(2014) work on five verb lemmas occurring with the 'intrusive *as*' construction has been the first major corpus-based investigation of the feature. Her findings present the construction not only as a genuine and widely spread pattern in IndE, even in the written mode, but indeed as being a truly pan-South Asian feature. Parallel to e.g. *known as*, verbs such as *name* and *deem* (among others) occur with the additional particle in all South Asian second language varieties of English:

(1) the new board declared him as the chairman (SAVE-IN_SM_2004-07-15.txt)
(2) they termed it as a 'tragic event' (SAVE-LK_DM_2005-02-28.txt)
(3) Among traditional plantation crops (tea, rubber and coconut) this was named as a "lazy man's crop" (LK_DN_2004-06-16.txt)

On the basis of data from the South Asian Varieties of English (SAVE) corpus (Bernaisch et al. 2011), the present study seeks to deepen our understanding of this feature's origin by taking the whole paradigm of verbs occurring in the complex-transitive complementation pattern into account. We expect 'intrusive *as*' to be either motivated by the Dravidian and Indo-Aryan quotative particle or to be an effect of 'nativized semantico-structural analogy' (Mukherjee 2007) that aims at regularization or a higher level of explicitness within the set of complex transitive verbs. Taking several context factors such as voice, object length and complement type into account, the study aims at providing a detailed complementational profile of all verb lemmas that allow the 'intrusive *as*' construction. If the contact hypothesis holds, the relevant constructions would be expected to mirror the usage of the quotative particle in the substrate languages.

In a secondary step, the analysis will be extended to a selection of learner corpora in order to examine potential similarities between the postcolonial contexts in South Asia and English-learning scenarios. If the feature is similarly frequent in both second-language varieties and learner varieties of English, as preliminary studies seem to indicate, this would lend support to the idea that the 'intrusive *as*' construction is not restricted to specific contexts but rather likely to be a universal of language acquisition.

References

Bernaisch, T., C. Koch, J. Mukherjee & M. Schilk. 2011. *Manual for the South Asian Varieties of English (SAVE) Corpus*. Giessen: Justus-Liebig-University, Department of English.

Goffin, R. C. 1934. Some notes on Indian English. *S.P.E. Tract No. XLI.* Oxford: Clarendon Press.

Lange, C. 2014. 'People call it as city of garden' – tracing the 'intrusive as'-construction in South Asian varieties of English. Paper presented at SALA 30, Central University Hyderabad, 6-8 Feb 2014.

Mukherjee, J. 2007. Steady states in the evolution of New Englishes: present-day Indian English as an equilibrium, *Journal of English Linguistics*, 35(2): 157-87.

Nihalani, Paroo, R. K. Tongue & P. Hosali. 1979. *Indian and British English: A Handbook of Usage and Pronunciation*. New Delhi: Oxford University Press.

Sedlatschek, A. 2009. *Contemporary Indian English: Variation and Change*. Amsterdam: Benjamins.

Yadurajan, K. S. 2001. *Current English: A Guide for the User of English in India.* New Delhi: Oxford University Press.

Editorial practice and the distinction between error and conventionalised innovation in New Englishes: A corpus-based investigation of Black South African English

Haidee Kruger (Macquarie University / North-West University)
Bertus van Rooy (North-West University)

Language change essentially depends on two processes: innovation and propagation (Croft 2000). Innovation is the process by which new form-function mappings are created by individuals, whereas propagation involves the selection of such mappings by means of social processes, leading to conventionalisation. These two dimensions, individual psycholinguistic innovation and social conventionalisation of innovative forms, form the basic dynamic of all language change, and may be seen to apply to not only native varieties of English but also to the indigenised non-native varieties of English, or New Englishes (see Schneider 2007).

Van Rooy (2011) points out that one of the key questions in understanding the development of the New Englishes is the distinction between error and conventionalised innovation. The language-contact environment in which the New Englishes develop offers diverse opportunities for innovations, ranging from performance errors to substrate interference at lexical, morphological and syntactic levels. The question arises what the social mechanisms are that allow these innovations, sometimes starting as errors or deviations from normative usage in the native variety, to become conventionalised features of the new variety. These mechanisms are complex, but Bambgose (1998) argues that one criterion for the distinction between error and innovation is acceptance by standardising authorities, such as the media and publishing houses. In this respect, editorial practice may provide an indication of the degree to which particular features are viewed as errors, or has become accepted features of the variety in question, since the "publishing industry itself, and the editorial profession, are not neutral parties in maintaining public awareness of usage sanctions… They have a gatekeeper role in enforcing selected usage practices…" (Peters 2006).

In this paper, we investigate this hypothesis using a parallel corpus of unedited and edited versions of texts produced by speakers of Black South African English. A comparative analysis of editorial changes (with particular attention to features associated with BSAE) is carried out, with the aim of investigating whether editorial intervention provides a measure to draw a distinction between error and conventionalised innovation, at the lexical and morphosyntactic levels. Given the fact that different registers are more or less receptive to language change, and that editorial practices differ for different text types (see Biber & Gray 2013; Hundt & Mair 1999), the corpus includes newswriting (as an example of a register more amenable to change) as well as academic writing (as an example of a more conservative register).

References

Bambgose, A. 1998. Torn between the norms: innovations in world Englishes. *World Englishes*, 17(1): 1-14.

Biber, D. & B. Gray. 2013. Being specific about historical change: the influence of sub-register. *Journal of English Linguistics,* 41(2): 104-34.

Croft, W 2000. *Explaining Language Change: An Evolutionary Approach*. London: Longman.
Hundt, M. & C. Mair. 1999. 'Agile' and 'uptight' genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics*, 4: 221-42.
Peters, P. 2006. English usage: prescription and description. In B. Arts & A. McMahon eds. *The Handbook of English Linguistics*, London: Blackwell. Blackwell Reference Online. http://www.blackwellreference.com/subscriber/tocnode.html?id=g9781405113823_chunk_g978140511382333.
Schneider, E. 2007. *Post-colonial Englishes: Varieties Around the World.* Cambridge: Cambridge University Press.
Van Rooy, B. 2011. A principled distinction between error and conventionalised innovation in African Englishes. In J. Mukherjee & M. Hundt eds. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*, Amsterdam: Benjamins, 191-209.

The fate of linguistic innovations in Jersey English

Anna Rosen (University of Freiburg)

Jersey English, a small and lesser known variety on the periphery of continental Europe, has emerged over the last centuries in a linguistic and cultural contact setting between Norman French, French and English. It can be described, following Mesthrie's (1992) terminology, as a nativized L2 variety, as the shift towards a monoglot English society is by now almost complete with only 2 to 3% of the population still speaking Norman French. Jersey's more remote geographical location in combination with diverse linguistic influences and a turn away from agriculture towards an international finance centre makes it an interesting test case for hypotheses in contact linguistics and makes matters of identity stand out. While the linguistic norm in Jersey very much leans towards an exonormative British English standard, local usage patterns shaped by language and (ongoing) dialect contact sometimes run counter to this.

Based on sociolinguistic interview and archive data, collected in 2008 and compiled to a 350,000-word corpus of spoken Jersey English, this paper offers a historical perspective on linguistic innovations in non-native Englishes: It traces the emergence and development of three characteristic features of Jersey English, existential *there's* with a time reference a Verb-*and*-Verb construction (where the second verb is invariantly in the infinitival form) and the pragmatic particle *eh* as illustrated in examples (1), (2) and (3).

(1) There's sixty years we're married.

(2) I went and buy some pansy plants.

(3) It's good for the cattle, eh?

The existence of all three features in Jersey English can be attributed to contact. Whereas the specific existential construction can be considered a transfer phenomenon, the coordinated verb structure seems to be the result of imperfect group learning and the particle *eh* of the language contact situation more generally (cf. Ramisch 1989; Barbé 1995; Jones 2001; Rosen 2014). The corpus data allow us to establish, in retrospect, that these linguistic innovations became conventionalized and widely accepted in the Jersey English speech community but

also to speculate on their future development. The current frequency distribution across age, social and linguistic background of Jersey English speakers suggest that *eh*, though socially stratified in its use, survives as a stereotypical and identity-constructing feature whereas both existential patterns as in (1) and V-*and*-V constructions as in (2) are declining and only used in very specific speaker networks. The paper therefore argues that two pressures act on processes of variation and change in Jersey: pride in local norms and economic pressure for a standard which is globally accepted. Metalinguistic discussions in the corpus and background information about the speakers support this line of argument and reveal a complex web of attitudes, feelings of identity and linguistic awareness issues. It thus can be shown which processes and factors are important in shaping the use and survival of such former innovations.

References

Barbé, P. 1995. Guernsey English: a syntax exile? *English World-Wide*, 16.1: 1-36.

Jones, M. C. 2001. *Jersey Norman French*: *A Linguistic Study of an Obsolescent Dialect.* Oxford: Blackwell.

Mesthrie, R. 1992. *English in Language Shift. The History, Structure and Sociolinguistics of South African Indian English*. Cambridge: Cambridge University Press.

Ramisch, H. 1989. *The Variation of English in Guernsey/Channel Islands.* Frankfurt/M.: Lang.

Rosen, A. 2014. *Grammatical Variation and Change in Jersey English*. Amsterdam: Benjamins.

Detecting innovations in a parsed corpus of learner English

Gerold Schneider (University of Zurich)
Gaëtanelle Gilquin (FNRS, Université catholique de Louvain)

The concept of linguistic innovation in English has so far mainly been limited to the description of native and indigenized varieties (ESL). In foreign varieties of English (EFL), on the other hand, non-standard forms are typically considered as errors. Such a treatment, however, (i) fails to acknowledge those cases when foreign learners intend to be creative, as underlined by Rimmer (2008), and (ii) misses commonalities between ESL and EFL. Recent corpus-based studies have provided preliminary evidence that some non-standard forms are shared by indigenized and foreign varieties of English. Nesselhauf (2009) has brought to light similarities in the way of new prepositional verbs like *comprise of* or *emphasize on*, while Gilquin (2011) has drawn parallels between phrasal verbs in ESL and EFL (see also Götz & Schilk 2011, Davydova 2012, Laporte 2012 and Deshors 2014, among others). Such commonalities challenge the idea of a clear dichotomy between innovations and errors, and encourage us to look for more similarities between ESL and EFL.

We present a data-driven method to detect potential innovations in EFL on a large scale, test it on verb-preposition structures, and describe similarities and differences between ESL and EFL. Relying on the whole of the *International Corpus of Learner English* (ICLE), which has been parsed with the probabilistic dependency parser Pro3Gres (Schneider 2008), we have automatically extracted potential innovations, defined here as patterns of overuse in

ICLE compared to a reference corpus, for which we use the *British National Corpus* (BNC). We measure overuse by means of various collocation measures such as O/E or T- score (e.g. Evert 2009). Our approach is related to Schneider & Zipp (2013), which allows us to conduct a detailed comparison with novel combinations of verbs and prepositions found in Schneider & Zipp (2013) for ESL, based on the *International Corpus of English* (ICE). We find both striking similarities (e.g. *discuss about*) and dissimilarities (e.g. *accuse for*, only distinctive for EFL).

The quantitative study is followed by a qualitative step, in which we aim to explain origins of non-native-like combinations in EFL (e.g. *viewed upon as*, probably built by analogy with *looked upon as*) and try to find criteria to determine what could be identified as actual innovations. We discuss total frequency, recurrence limited to learners from the same L1, which could point to L1 transfer innovations, and recurrence across different L1s, which could point to psycholinguistically based innovations that are the result of, e.g., processing load or semantic explicitness.

References

Davydova, J. 2012. Englishes in the outer and expanding circles: a comparative study. *World Englishes*, 31: 366-85.

Deshors, S. C. 2014. A case for a unified treatment of EFL and ESL: a multifactorial approach. *English World-Wide*, 35(3): 277-305.

Evert, S. 2009. Corpora and collocations. In A. Lüdeling & M. Kytö eds. *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 1212-48.

Gilquin, G. 2011. Corpus linguistics to bridge the gap between World Englishes and Learner Englishes. In L. Ruiz Miyares & M. R. Álvarez Silva eds. *Comunicación social en el siglo XXI, Vol. II*. Santiago de Cuba: Centro de Lingüística Aplicada, 638-42.

Götz, S. & M. Schilk 2011. Formulaic sequences in spoken ENL, ESL and EFL: focus on British English, Indian English and learner English of advanced German learners. In J. Mukherjee & M. Hundt eds. *Exploring Second- Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: Benjamins, 79-100.

Laporte, S. 2012. Mind the gap! Bridge between World Englishes and Learner Englishes in the making. *English Text Construction*, 5: 265-92.

Nesselhauf, N. 2009. Co-selection phenomena across New Englishes: parallels (and differences) to foreign learner varieties. *English World-Wide*, 30(1): 1- 25.

Rimmer, W. 2008. Grammatical creativity in learner corpora. *Humanising Language Teaching*, 10(1). http://www.hltmag.co.uk/jan08/idea.htm.

Schneider, G. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. PhD Thesis. Institute of Computational Linguistics, University of Zurich.

Schneider, G. & L. Zipp 2013. Discovering new verb-preposition combinations in New Englishes. *Studies in Variation, Contacts and Change in English*, 13. http://www.helsinki.fi/varieng/series/volumes/13/schneider_zipp

Workshop III: Words, words, words – Lexis, corpora and contrastive analysis

Convenors: Signe Oksefjell Ebeling (University of Oslo) & Thomas Egan (Hedmark University College)

Comparing the language of children's literature

Anna Čermáková & Lucie Chlumská (Charles University, Czech Republic)

Children's literature and its language have been researched marginally despite the important role it plays socially and educationally. What constitutes children's literature is a debatable issue; however, the genre is mostly defined by extralinguistic criteria (Knowles & Malmkjær 1996). Genre categorisation creates both reader and writer expectations and these are then further carried over to translations including the expectations in the target language literary tradition that may be somewhat different. Expectations related to the translation of children's literature are particularly high especially for readability and naturalness. Potential manipulation of ideological elements (Puurtinen 1998:1) is of concern for this genre as well because it displays very unbalanced power relationship between the authors and readers (e.g. Hunt 1992, Knowles & Malmkjær 1996). This issue is particularly accented for translations into smaller languages because the volume of translations on the book market for children is fairly high.

The wider aim of our study is to explore cross-linguistically (between English and Czech) linguistic and discourse features of children's literature. The study aims to be corpus-driven (Tognini Bonelli 2001) and one of the approaches to be taken is through n-grams. N-grams have been applied in corpus linguistics extensively as one of the ways to explore the operating idiom principle (Sinclair 1991). The particular aim of this study is to establish through n-grams frequent patterns as pointers to potential units of meaning that may be comparable in the two languages. The study is methodologically inspired by Ebeling & Oksefjell Ebeling (2013), Tognini Bonelli (2002) and Mahlberg (2007).

Working with two typologically different languages shows that the n-gram method is potentially challenging. The suitable length of the n-grams to be further explored varies among researchers. Ebeling & Oksefjell Ebeling (2013) use 3-grams in their general contrastive study of phraseology between English and Norwegian (typologically close languages) while Mahlberg (2012) finds most satisfying 5-grams to explore Dickens's fiction. Due to the inflectional nature of Czech, it is impossible to decide on the corresponding length for English 3-grams and 5-grams respectively. Therefore, we need to establish the suitable length of n-gram for Czech independently, i.e. not through translational equivalence in the parallel corpus. We will explore various lengths of n-grams in comparable corpora: a corpus of English children's literature, a corpus of original Czech children's

literature and a corpus of translated children's literature while validating the results in a parallel translation corpus (InterCorp).

References

Ebeling, J., & S. Oksefjell Ebeling. 2013. *Patterns in Contrast.* Amsterdam: Benjamins.

Hunt, P. ed. 1992. *Literature for Children: Contemporary Criticism.* London & New York: Routledge.

Knowles, M. & K. Malmkjær. 1996. *Language and Control in Children's Literature.* London: Routledge.

Mahlberg, M. 2012. *Corpus Stylistics and Dickens's Fiction*. London: Routledge.

Mahlberg, M. 2007. Corpora and translation studies: textual functions of lexis in *Bleak House* and in a translation of the novel into German. In V. Intouti, G. Todisco & M. Gatto eds. *La Traduzione. Lo Stato dell'Arte. Translation. The State of the Art.* Ravenna: Longo Angelo, 116-135.

Tognini Bonelli, E. 2001. *Corpus Lingustics at Work.* Amsterdam: Benjamins.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Corpus resources

SYN

Czech National Corpus – SYN. Institute of the Czech National Corpus, Praha. 13.01.2015: http://www.korpus.cz.

InterCorp (version 6)

Czech National Corpus – InterCorp, Institute of the Czech National Corpus, Praha. http://www.korpus.cz.

Novice translators' rendering of supplementive *-ing* clauses

Hildegunn Dirdal (University of Oslo)

English *-ing* clauses are both frequent and multifunctional. Since Norwegian does not have an equivalent structure, translators are forced to find other solutions. While several studies have investigated professional Norwegian translators' rendering of *-ing* clauses (e.g. Behrens 1998, Behrens and Fabricius-Hansen 2005, Fuhre 2010, Smith 2004), there is a lack of research on novice translators. The aim of this study is to investigate the challenges translation students face, with the hope that the findings can lead to improved teaching.

The study focuses on subjectless supplementive *-ing* clauses, whose semantic relationship to the matrix clause is not specified and is often ambiguous. Quirk et al. (1985: 1123) compare their versatility to that of *and*, and coordination is in fact the most frequently used structure in Norwegian translation of supplementive *-ing* clauses (Fuhre 2010). However, the likelihood of coordination being chosen depends on the semantic relation, and not all relations can be expressed in this way (Dirdal, submitted). Professional translators also show a tendency to disfavour this solution when coordination is already present somewhere else in the sentence (Dirdal, submitted). The present study investigates whether student translators have problems with the rendering of specific semantic relations and whether they take into account the structure of the rest of the sentence. More specifically, the following hypothesis was formulated:

Student translators will overuse coordination, choosing it in cases where

(a) it cannot indicate the semantic relation that holds between the *-ing* clause and its matrix in the source text

(b) the presence of other coordination makes the sequence awkward

Data was collected from a corpus of 24 master theses in literary translation from English to Norwegian. All sentences with *-ing* clauses were extracted from the source texts and matched with their translations in the target texts. The subjectless supplementive clauses were selected for analysis related to the hypothesis above.

Preliminary results lend support to the hypothesis. There is a certain amount of overuse of coordination that is problematic either because coordination does not indicate the same semantic relation between the clauses as the *-ing* clause has to its matrix in the source text, or because of an awkward string of coordinations in the sentence. The latter is related to a broader problem for the students: not primarily the choice of structure to render the *-ing* clause, but the integration of this structure into a more complex sentence.

References

Behrens, B. 1998. *Contrastive Discourse: An Interlingual Approach to the Interpretation and Translation of Free ING-participial Adjuncts*. PhD, University of Oslo.

Behrens, B. & C. Fabricius-Hansen. 2005. The relation Accompanying Circumstance across languages. Conflict between linguistic expression and discourse subordination? *SPRIKreports* 32. Oslo: University of Oslo. http://www.hf.uio.no/ilos/forskning/prosjekter/sprik/pdf/bb/Sprik-Report32-bb-cfh.pdf. Accessed 08.10.2014.

Dirdal, H. [submitted] Factors influencing the translation of *-ing* clauses: Semantic role, context and a translator's individual style. *ICAME 35 Proceedings*.

Fuhre, P. 2010. The English -ing Participial Free Adjunct in Original and Translated Fiction: An English-Norwegian Parallel Corpus Study. Master's thesis, University of Oslo.

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

Smith, M.-B. Marthinsen. 2004. *Initial* -ing *in English and their Translation into Norwegian*. Master's thesis, University of Oslo.

A corpus-based contrastive analysis of English and French lexical bundles across genres

Natalia Grabar (STL UMR8163 CNRS, Université de Lille 3)
Marie-Aude Lefer (Marie Haps School for Translators and Interpreters, Brussels)

This presentation reports on a large-scale analysis of English and French lexical bundles (also called recurrent word combinations) in five genres. Phraseological units, especially lexical bundles, have been largely under-researched in corpus-based contrastive studies so far (cf. Ebeling & Ebeling 2013). However, bundles are definitely worth investigating, as they can help uncover metadiscursive and rhetorical cross-linguistic contrasts that could not be revealed otherwise.

Looking at parliamentary debates and newspaper editorials, Granger (2014) found that lexical bundles are used more pervasively in French argumentative discourse than in English. This can be attributed to a systemic difference between the two languages: French has often been said to rely heavily on rhetorical markers, to be explicitly emphatic and, more generally, to be more verbose than English (see e.g. Vinay & Darbelnet 1995). Drawing on insights from recent cross-register contrastive studies (Hansen-Schirra et al. 2012, Neumann 2013, Lefer & Vogeleer 2014), we wish to examine the genre-sensitivity of lexical bundles in English and French. One of our starting-point hypotheses is that, in view of the pervasiveness of bundles in French, genres should share more bundles in French than in English.

The contrastive analysis relies on comparable data extracted from four corpora, representing five genres: Europarl (transcripts of parliamentary debates; Koehn 2005, Cartoni & Meyer 2012), KIAP (research articles in medicine, economics and linguistics; Fløttum et al. 2006), Multed (editorials) and PLECI (news and fiction). The presentation focuses on two case studies: Case Study 1 looks at all five genres, relying on 350,000-word corpora per genre and per language, while Case Study 2 is restricted to four genres (all genres mentioned above, except fiction), making use of one-million-word corpora per genre and per language. In the two case studies, we examine trigrams (3-word lexical bundles) only. All trigrams were automatically extracted, of which the most frequent 1% trigram types in each corpus were kept for further analysis. This corresponds to ca. 7,000 bundles per language in Case Study 1 and ca. 15,000 bundles per language in Case Study 2. These were then automatically classified as being genre-specific (i.e. attested in one genre only) or shared by genres (i.e. attested in 2, 3, 4 or 5 genres), representing a continuum from high genre-sensitivity (or specificity) to high genre-insensitivity.

Contrary to our initial expectations, preliminary quantitative results suggest that English and French are quite similar as regards the distribution of bundles across genres: in Case Study 1, ca. 75% of the top-1% trigrams are genre-specific in each language (English examples include *however I believe*$_{\text{Europarl}}$, *his voice was*$_{\text{PLECIfiction}}$, *empirical analysis of*$_{\text{KIAP}}$), while 1%-2% are genre-insensitive, i.e. shared by all genres (e.g. *some sort of*, *in the process*, *than that of*, *on the contrary*, *no reason to*). Similar results are obtained in Case Study 2.

In our presentation we will examine the cross-genre similarities and differences in the two languages, so as to uncover some of the typical phraseological features of the genres and language systems under investigation. We will also try to characterize the genre-specific and genre-insensitive trigrams, relying, among other things, on the distinction between referential bundles, discourse organizers and stance markers (Biber *et al*. 2004). The presentation will end with some of the implications of our study for corpus-based cross-linguistic phraseological research.

References

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

Cartoni, B. & T. Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *8th International Conference on Language Resources and Evaluation* (LREC).
Ebeling, J. & S. Oksefjell Ebeling. 2013. *Patterns in Contrast*. Amsterdam: Benjamins.
Fløttum, K., T. Dahl & T. Kinn. 2006. *Academic Voices – Across Languages and Disciplines*. Amsterdam: Benjamins.
Granger, S. 2014. A lexical bundle approach to comparing languages: stems in English and French. *Languages in Contrast*, 14(1): 58-72.
Hansen-Schirra, S., S. Neumann & E. Steiner. 2012. *Cross-Linguistic Corpora for the Study of Translations - Insights from the Language Pair English-German*. Berlin: de Gruyter Mouton.
Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. In *MT Summit X*, 79-86.
Lefer, M.-A. & S. Vogeleer eds. 2014. *Genre- and Register-Related Discourse Features in Contrast*. Special issue of *Languages in Contrast*, 14(1).
Neumann, S. 2013. *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin: de Gruyter Mouton.
Vinay, J.-P. & Darbelnet, J. 1995. [1958]. *Comparative Stylistics of French and English. A Methodology for Translation*. Amsterdam: Benjamins.

Lexical patterns of place in English and Norwegian

Hilde Hasselgård (University of Oslo)

This study aims to explore the semantic field of place in two ways. It starts by investigating the cognates *place* and *plass* in the fiction part of the English-Norwegian Parallel Corpus (ENPC) to survey their uses and translation correspondences (cf. Johansson 2007: 92). Because of its frequency both in Norwegian originals and as a translation of *place* the lexeme *sted* is also included in the analysis. The initial research question is thus:

- What do correspondence patterns reveal about similarities and differences between English *place* and Norwegian *plass* and *sted*?

A preliminary investigation shows that *place* is translated by *sted* approximately 41% of the time; *plass* accounts for about 11%. About 13% are untranslated. None of the remaining correspondences are particularly frequent, but many specify a type of *place*, e.g. *hus* ('house'), *område* ('area'), and some are place adverbs, e.g. *her* ('here'). *Plass* is translated as *place* about 40% of the time, while *room* accounts for 12%. Other recurrent translations highlight the polysemy of *plass* (e.g. *seat, space, square, job*). In translations of *sted*, *place* dominates again (32%), but a striking share (23%) comprises indefinite adverbs such as *somewhere*. Translation patterns thus suggest that *place* and *sted* have more general meanings than *plass*.

The lexemes under study are sometimes part of set phrases, e.g. *in the first place*, *plass til* ('place [room] for'). Some of these are only loosely associated with location, e.g. *i stedet* ('in stead'), regularly corresponding to the cognate *instead*, or *finne sted* ('find place'), which usually corresponds to *take place*.

In a next step I will thus investigate the lexical surroundings of *place*, *plass* and *sted* and their most frequent correspondences to answer this question:

- What can recurrent word combinations reveal about the ways in which place is referred to in English and Norwegian?

I will use AntConc to identify recurrent n-grams involving *place, space, room, plass, sted* and *rom*. Only n-grams occurring at least four times and in at least two texts will be considered. This frequency criterion excludes 5-grams; furthermore, a selection will be made of 2-4-grams that show some degree of semantic unity and are phraseologically interesting (see Ebeling & Ebeling 2013: 69; Altenberg 1998: 102), either in themselves or because of their translation patterns. Preliminary investigations suggest that the lexemes overlap very little as regards their lexical surroundings. The combined investigation of translation patterns and phraseology is expected to illuminate the division of labour between these near-synonyms denoting the concept of place in both languages.

References

Altenberg, B. 1998. On the phraseology of spoken English: the evidence of recurrent word-combinations. In A.P. Cowie ed. *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 101-122.

Ebeling, J. & S. Oksefjell Ebeling. 2013. *Patterns in Contrast*. Amsterdam: Benjamins.

Johansson, S. 2007. *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam: Benjamins.

"One I don't hate as much as the others": Expressing love in Finnish and English

Mikko Höglund (Stockholm University)

Stereotypically, Finns are seen as serious, quiet, polite, straightforward and honest people (Vihakara 2006, Kirra 1999, Lehtonen & Sajavaara 1985). The combination of these qualities, at least to outsiders, may also convey a sense of distance and indifference. In fact, this is a stereotype that is kept alive by Finns themselves as well; it is not just something that is cast upon them by other people. For instance, Uusiautti and Määttä (2012) found out that what Finns find important in love is not only the sentiment but also the knowledge and skills that are required, and that love can be learned and controlled. The claim of the Finns' cold-heartedness is often accompanied by the argument that Finns do not use the word *rakkaus* 'love' very frequently, and that the word does not fit the sentiment it expresses. This notion is clearly expressed in a popular Finnish song "Rakkaus on ruma sana" ('Love is an ugly word', Alanko 1998), in which the word is characterized as "the rapist of poems" and "a monster of a word".

The aim of this paper is to use corpus data to compare the use of the words conveying the meaning 'love' (noun/verb/adjective) in Finnish and English. Whereas the situation in Finnish is as described above, English uses the word *love* much more readily, and especially

in American English the word is very prominent and used in a variety of situations (e.g. Werner 2012, Tissari 2004). The paper presents a contrastive study that looks at the issue from a quantitative perspective comparing the frequencies of 'love' in corpus data as well as the contexts in which 'love' is used. First, two independent corpora, COCA and the Language Bank of Finland, are searched for the relevant terms, and the results in different text types are compared. Second, the Tampere Bilingual Corpus of Finnish and English (TamBiC), which consists of original English texts with their Finnish translations and vice versa, will be searched in order to see how *love* is translated into Finnish.

The pilot studies in COCA and the Language Bank of Finland show that *love* is used more than *rakkaus*, especially in magazines. The same trend should be seen in TamBiC as well, and that a great deal of the instances of *love* in the original English texts are not translated using the word *rakkaus*. It is further hypothesized that the lack of *rakkaus* is not due to the lack of the concept itself but that the concept is expressed in other ways and using different vocabulary in the Finnish translations.

References

Alanko, I. 1998. Rakkaus on ruma sana. On *Pulu* [CD]. Poko Rekords.

Kirra, K-M. 1999. *The Types of Problematic Phenomena Perceived by Finns in Their Communication with Non-Finns: A Study of Critical Incidents*. Unpublished M.A. Thesis. Jyväskylä: University of Jyväskylä.

Lehtonen, J. & K. Sajavaara. 1985. The Silent Finn. In D. Tannen & M. Saville-Troike eds. *Perspectives on Silence*. Norwood, NJ: Ablex Publ. Corp,. 193-201.

Uusiautti, S. & K. Määttä. 2012. The ability to love – a virtue-based approach. *British Journal of Educational Research, 2*(1): 1-19.

Tissari, H. 2004. *Like* like *love*: comparing two modern English words diachronically. In C. J. Kay, C. A. Hough, & I. Wotherspoon eds. *New Perspectives on English Historical Linguistics. Volume II: Lexis and Transmission.* Amsterdam: Benjamins, 235-49.

Vihakara, A. 2006. *Patience and Understanding. A Narrative Approach to Managerial Communication in a Sino-Finnish Joint Venture*. Turku: Turku School of Economics and Business Administration.

Werner, V. 2012. Love is all around: a corpus-based study of pop lyrics. *Corpora, 7*(1): 19-50.

English hyphenated premodifiers in German and Swedish translations: A cutting-edge state-of-the-art study

Magnus Levin & Jenny Ström Herold (Linnaeus University)

This study stems from our work training translators where we have noticed that trainee translators struggle with English hyphenated premodifiers. Such premodifiers come in a variety of different forms, e.g., N + *ed*-participle (*pig-headed losers*), adjective + *ing*-participle (*a tight-fitting beret*), NPs (*the end-of-term reports*) and verb phrases (*a go-along-and-enjoy-yourselves gesture*) (for an overview, see Biber et al. 1999: 534–5). As indicated in these examples, hyphens are used both with highly lexicalized premodifiers (Arnaud et al. 2008: 116) and ad hoc constructions. The aim is to investigate how professional translators translate these construction types into German and Swedish in the Oslo Multilingual Corpus,

focusing on how the structural means of the two target languages affect the choices made. The results will also help to improve teaching materials for trainee translators, by providing an overview of the strategies used by professionals.

Previous findings suggest that premodification is more common in German than in Swedish source texts, which favour postmodification (cf. Fleischer & Barz 1995: 320–31; Teleman et al. 1999: III: 71–84). It can therefore be assumed that translations into these languages also have different preferences.

Our data show that different construction types are connected to different types of translation alternatives, and there are some indications of target-language-specific preferences. For example, *ed*-participles are generally rendered as similar adjectives in the translations (*liver-coloured > leberfarbene/leverfärgad*), and relative clauses are more common in Swedish translations (*a market-analysis firm > ett företag som gjorde marknadsanalyser*) which confirms the observation that Swedish is more prone to postmodification. In the German translations, on the other hand, complex premodifications are more often rendered as extended participial premodifiers (*sea-washed stone > vom Meer glattgeschliffenen Stein*). Other frequent translation strategies involve compounding and prepositional phrases. These premodifiers often lack lexicalized equivalents, and they are often omitted, restructured or rendered word-for-word.

Our results indicate that there is a relatively low degree of correspondence between the structures chosen in the target languages, at least regarding the more lexicalized instances. This suggests that the structural means of the individual languages affect the strategies used by either allowing or forcing translators into making different choices. The degree of lexicalization is a key factor when translating more freely or word-for-word.

References

Arnaud, P. J. L., E. Ferragne, D. M. Lewis & F. Maniez. 2008. Adjective + Noun sequences in attributive or NP-final positions: observations on lexicalization. In S. Granger & F. Meunier eds. *Phraseology: An Interdisciplinary Perspective*. Amsterdam: Benjamins, 111-25.

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

Fleischer, W. & I. Barz. 2012. *Wortbildung der deutschen Gegenwartssprache*. Berlin: De Gruyter.

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

Non-prepositional English equivalents of Czech prepositional phrases: From function words to functional sentence perspective

Markéta Malá (Charles University in Prague)

In corpus-linguistic studies, preposition-based sequences have been found, for instance, to characterize text-types (Scott & Tribble 2006, Hunston 2008, Groom 2010) or differentiate cross-linguistically between concepts encoded by prepositions (Egan 2013). In this study, the

translation paradigms of prepositions are shown to be indicative of typological differences between languages at the level of phrases and clauses. English divergent translation correspondences of Czech prepositions point out some consequences of the predominantly analytic character of English, highlighted through the comparison with the synthetic inflectional Czech language. The study deals with English translation counterparts of four most common Czech prepositions – *v/ve, na, s/se, z/ze* ('in, on, with, from'). It focuses on the overt non-prepositional correspondences, which constitute 20.2%, 17.7%, 27.0% and 19.2% of counterparts of *v/ve na, s/se* and *z/ze*, respectively.

The English non-prepositional counterparts include formally and functionally diverse constructions, such as adverbs in the function of the adverbial, or (non-)finite clauses. While some of these correspondences are conditioned lexically (e.g. the translation of a Czech prepositional object by an English direct one: *znovu se scházet se mnou* 'meet with me' – *to meet me again*), some of the most frequent types appear to be quite systematic and conditioned typologically:

Czech postmodifying prepositional phrases translated by English participial postmodification, showing the semantically reduced near-prepositional status of the English participle, e.g. *dřevěné regály s jejími výtvory* ('shelves with her work') – *wooden shelves holding her work*.

The Czech prepositional phrase used as a postmodifier where English relies on premodification by a noun, e.g. *pouzdro s pistolí* ('a holster with a pistol') – *his pistol holster*, *sukni z jemné kůže* ('skirt from soft leather') – *her soft leather skirt*. In English the modifying function of the noun is signalled by the position with respect to the head noun; in Czech premodifiers are typically adjectives whose relationship to the head noun is indicated by inflectional suffixes. If the adjective cannot be used, Czech resorts to prepositional postmodification.

The consequences of English grammatical fixed word order are particularly prominent where the Czech adverbial prepositional phrase is paralleled by the English subject noun phrase, e.g. *ve městě bylo pět hotelů* ('in the town were five hotels') – *the town had several hotels*. The clause-initial position of the subject in English coincides with the unmarked position of the theme, i.e. an element which can convey information on the setting or circumstances of the content of the clause (Dušková 2002). Some English subjects therefore assume 'adverbial' semantic roles. In Czech, these roles are performed by prepositional phrases, with the subject (closely tied with the semantic role of the agent) occurring farther in the clause.

References

InterCorp: Český národní korpus – InterCorp. Ústav Českého národního korpusu FF UK, Praha. http://www.korpus.cz.

Dušková, L. 2002. Constancy of syntactic function across languages: In J. Hladký ed. *Language and Function: To the Memory of Jan Firbas*. Amsterdam: Benjamins. 127-45.

Egan, T. 2013. Between and through revisited. In M. Huber & J. Mukherjee eds. *Studies in Variation, Contacts and Change in English*. Vol 13. http://www.helsinki.fi/varieng/series/volumes/13/egan/.

Groom, N. 2010. Closed-class keywords and corpus-driven discourse analysis. In M. Bondi & M. Scott eds. *Keyness in Texts*. Amsterdam: Benjamins. 59-78.
Hunston, S. 2008. Starting with the small words. *International Journal of Corpus Linguistics*, 13(3): 271-95.
Klégr, A. et al. 2012. *Anglické ekvivalenty nejfrekventovanějších českých předložek*. Praha: Karolinum.
Scott, M. & C. Tribble 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: Benjamins.

Meaning shifts and use of collocations: A study comparing interpreting and translation into English and Italian

Maja Miličević (University of Belgrade)
Silvia Bernardini (University of Bologna)
Adriano Ferraresi (University of Bologna)

Within the wider contrastive paradigm, a methodological approach is emerging in corpus-based translation/interpreting studies, that relies on bilingual parallel corpora of interpreted and translated texts. These *intermodal* corpora are a source of valuable authentic data about task effects on the use of phraseology, since simultaneous interpreting involves online processing under time constraints while translation does not.

One of the few plurilingual intermodal corpora available is EPTIC, the *European Parliament Translation and Interpreting Corpus* (Bernardini et al provisionally accepted). EPTIC features transcripts of interpreted speeches and their sources, aligned to each other and to the corresponding translated versions and respective sources, in Italian and English. The corpus is part-of-speech tagged, lemmatised and indexed with the Corpus WorkBench. EPTIC currently contains 568 texts and about 250,000 words altogether.

Relying on EPTIC, we analyse bidirectionally the choices made by translators and interpreters that result in the presence of collocations in the target texts. Collocation candidates based on pre-selected part-of-speech patterns are extracted from the target language sub-corpora. To evaluate their collocation status, frequency data are obtained from itWaC and ukWaC (Baroni et al 2009) and used to calculate lexical association scores (Mutual Information and *t*-score); 150 collocations per sub-corpus are randomly extracted for manual analysis from among those with association scores above the EPTIC median for the relevant language and frequency>2. Following up on previous work in which we checked if the presence of a collocation in the target text corresponds to the presence of a collocation in the source text (Ferraresi et al submitted), in this paper we focus on meaning shifts (or lack thereof) occurring between the source text and target text fragment.

The target text collocation status (same/different meaning) is used as a binary outcome variable in a mixed-effects logistic regression model with language direction, mediation mode and association measure status as categorical predictors (fixed effects), controlling for possible influences of individual texts, part-of-speech patterns and collocations (random effects); the analysis is conducted using the *R* package *lme4* (Bates 2005).

Language direction, mediation mode and their interaction are found to contribute significantly to meaning maintenance/shift in target texts, while no effect is found for association measure status. Texts interpreted into English contain more meaning shifts than the corresponding translated texts, while no clear pattern emerges for mediated Italian; overall, the English texts contain more shifts than the Italian ones (Figure 1). We conclude that interpreters are more likely to perform meaning shifts, which might indicate that the preference for use of a collocation is routinized rather than conscious; however, the shifts also seem to be dependent on the language and mediation direction.

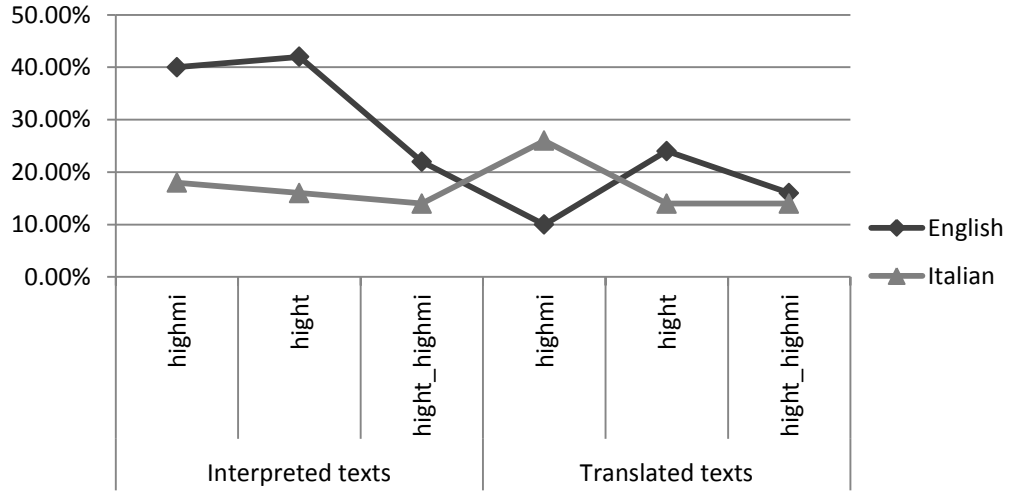


Figure 1. Percentages of meaning shifts by target language, mediation mode and association measures.

References

Baroni, M., S. Bernardini, A. Ferraresi, & E. Zanchetta. 2009. The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43, 3: 209-26.

Bates, D. 2005. Fitting linear models in R: using the lme4 package. *R News* 5: 27-30.

Bernardini, S., A. Ferraresi, & M. Miličević. [provisionally accepted]. From EPIC to EPTIC – exploring simplification in interpreting and translation from an intermodal perspective. *Target.*

Ferraresi, A., S. Bernardini, & M. Miličević. [submitted]. Collocations across languages: evidence from interpreting and translation.

A corpus-based analysis of genre-specific words: Minutes in English and Spanish

Rosa Rabadán (University of León, Spain)
Isabel Pizarro (University of Valladolid)
Marlén Izquierdo (University of the Basque Country EHU/UPV, Spain)

Meeting minutes are highly structured texts (Swales 1990) and show a significant degree of intertextuality (Bhatia 2004). English and Spanish minutes both present two vocabulary sets, one that codifies the 'field' (Halliday 2002 [1977]) and belongs in a given content area, and another that codifies the discursive practices of the genre 'minutes'. There are, however, significant differences in the way these discursive practices are codified within either language.

Previous work on minutes (Rabadán in press) highlights the existence of conventionalized language routines that tend to be associated with particular moves (Biber, Connor and Upton 2007).

This paper sets out to explore i) which words, and multi-word combinations (Gries 2008) can be identified as genre-and-move specific, and ii) what correspondences can be identified across languages.

Our study draws on an English-Spanish comparable corpus, C-GARE (Comparable corpus of meeting minutes in its Spanish acronym), an annotated corpus containing 100 texts in each language, featuring 139,919 words in English and 174,347 words in Spanish http://contraste2.unileon.es/web/en/corpus0_GARE.html tagged on the rhetorical level by means of the ACTRES text tagger. In addition, a comparable corpus browser (http://contraste2.unileon.es/web/en/tools.html) with a basic statistic feature has been used to obtain word list frequencies and multi-word combination patterns of up to four-grams within rhetorical moves in each language (Forchini and Murphy 2010). The identification and analysis of lexical patterns in contrast was based on Ebeling and Ebeling (2013), while the description of the grammatical category of the gram end-words and of the structural patterns, was based on Nesi and Basturkmen (2009). We explored the keyness of the genre through keywords and key-keywords (Scott 2010) to reveal the aboutness of each move and of the genre as a whole.

Empirical findings show that for each rhetorical move, irrespective of text 'field', a number of recurrent lexical patterns have become readily associated in each of the languages Since word choice is determined by genre-bound expectations and by the context, selections across languages are not obvious and correspondences show different structural patterns, word endings, and number of grams.

Results include the identification of genre-specific conventionalized lexis for each of the moves in both English and Spanish. These lexis-move associations can be successfully used in FL training and/or in writing aids (http://contraste2.unileon.es/apps/demos/gare/) to help L1 and L2 speakers of any of the two languages to produce effective texts when writing minutes (Bowker 2012).

References

Bhatia, V. K. 2004. *Worlds of Written Discourse. A Genre-Based View*. London: Continuum.

Biber, D., U. Connor & T. A. Upton eds. 2007. *Discourse on the Move. Using Corpus Analysis to Describe Discourse Structure.* Amsterdam: Benjamins.

Bowker, J. 2012. From 'communities of practice' to 'communities of learning': interdiscursivity in changing corporate priorities. In P. Gillaerts, E. de Groot, S. Dieltjens, P. Heynderickx & G. Jacobs eds. *Researching Discourse in Business Genres.* Bern: Peter Lang, 115-38.

Ebeling, J. & S. Oksefjell Ebeling. 2013. *Patterns in Contrast*. Amsterdam: Benjamins.

Forchini, P. & A. Murphy. 2010. N-grams in comparable specialized corpora: perspectives on phraseology, translation, and pedagogy. In U. Römer & R. Schulze eds. *Patterns, Meaningful Units and Specialized Discourses.* Amsterdam: Benjamins, 87-104.

Gries, S. Th. 2008. Phraseology and linguistic theory: a brief survey. In S. Granger & F. Meunier eds, *Phraseology: an Interdisciplinary Perspective*, Amsterdam: Benjamins; 3-25.

Halliday, M.A.K . 2002 [1977]. Text as semantic choice in social contexts. In J. J. Webster ed. *Linguistic Studies of Text and Discourse*. Vol 2; Collected Works of M. A. K. Halliday. London: Continuum, 23–81.

Rabadán, R. (in press). Proposals in meeting minutes: an English-Spanish corpus-based study. *Languages in Contrast*.

Nesi, H, & H. Basturkmen. 2009. Lexical bundles in discourse signaling in academic lectures. In J. Flowerdew & M. Mahlberg eds. *Lexical Cohesion and Corpus Linguistics,* Amsterdam: Benjamins, 23-44.

Scott, M. 2010. Problems in investigating keyness, or clearing the undergrowth and marking out trails… In M. Scott *Keyness in Texts,* Amsterdam: Benjamins, 43-58.

Swales, J. 1990. *Genre Analysis*. Cambridge: CUP.

Lexical variation in connector use: Do English and Norwegian argumentative texts differ?

Sylvi Rørvik (Hedmark University College)

There have been a number of cross-linguistic studies of connectors, defined as conjunctions or conjunct adverbials expressing logical relations between sentences. In comparison with English, Spanish has been shown to be less explicit, i.e. favoring a lower frequency of connectors (Dueñas 2009), whereas German, Swedish, and Danish, have been shown to be more explicit (Altenberg 1999, 2007; Shaw 2004; Kunz & Lapshinova-Kulinski 2014). Norwegian is closely related to Swedish and Danish, so one might expect similar results from a contrastive analysis of English and Norwegian, but no large-scale empirical study of connector use in Norwegian has yet been carried out. The present paper is an attempt to fill this gap.

The material for this study consists of argumentative newspaper texts in English and Norwegian (classified as 'expert' material), and argumentative student essays in English and Norwegian (classified as 'novice' material). The novice material has been included for pedagogical reasons, as a comparison of novice and expert writers within each language sheds light on the learning-to-write process, and allows us to identify potential developmental factors influencing connector use. In total, the material comprises some 347,000 words, divided between the 100 texts in each subcorpus. All connectors in the material have been extracted manually and assigned to one of Halliday's meaning categories (2004: 542-543):

apposition, clarification, addition, variation, spatio-temporal, manner, causal-conditional, and matter.

The paper attempts to answer the following questions:

1. Which of the two languages has the highest frequency of connectors, and does higher overall frequency correlate with greater lexical variation?
2. Do the languages differ in the lexical variation within the various meaning categories? If so, how?
3. Are there any cross-linguistic correspondences in the lexical items that are used most frequently within the various meaning categories?
4. Do novice and expert writers of English and Norwegian exhibit the same tendencies with respect to connector use and lexical variation?

Preliminary results indicate that Norwegian writers, both novice and expert, use more connectors than English writers, but that English has greater lexical variation. For all subcorpora a relatively short list of lexical items accounts for the vast majority of connectors. In addition, the number of low-frequency items varies between meaning types and from subcorpus to subcorpus.

References

Altenberg, B. 1999. Adverbial connectors in English and Swedish: semantic and lexical correspondences. In H. Hasselgård, & S. Oksefjell eds. *Out of Corpora. Studies in Honour of Stig Johansson.* Amsterdam: Rodopi, 249-68.

Altenberg, B. 2007. The correspondence of resultive connectors in English and Swedish. *Nordic Journal of English Studies*, 6:1.

Dueñas, P.M. 2009. Logical markers in L1 (Spanish and English) and L2 (English) business research articles. *English Text Construction.* 2(2): 246-64.

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

Kunz, K. & E. Lapshinova-Koltunski. 2014. Cohesive conjunctions in English and German: systemic contrasts and textual differences. In L. Vandelanotte, K. Davidse, K. Gentens & D. Kimps eds. *Recent Advances in Corpus Linguistics. Developing and Exploiting Corpora.* Amsterdam: Rodopi, 229-62.

Shaw, P. 2004. Sentence openings in academic economics articles in English and Danish. *Nordic Journal of English Studies*, 3(2): 67-84.

Reporting verbs in research writing: Cultural, disciplinary and genre perspectives

Jolanta Šinkūnienė (Vilnius University)

The increasing use of English as an academic lingua franca has been contributing to the growing body of research on academic rhetoric from both cross-disciplinary and cross-linguistic perspectives. Ken Hyland describes language variation in different disciplines as being "now one of the more fruitful lines of research" (Hyland 2011: 178), while the

comparison of how academic discourse is crafted in English versus a variety of other cultures has become a trendy topic in applied linguistics and EAP. A lot of research on academic rhetoric has focused on structural choices authors make in particular genres as well as on the use of metadiscourse markers. While literature abounds in studies on hedges, boosters, moves and steps in research articles, some of the areas of academic rhetoric are less researched, especially from a cross-linguistic perspective. A case in point is the use of reporting verbs in research writing. The choice of the verb to introduce other scholars' ideas is linked with author stance (cf. Hyland 2002) and evaluation (Thomson & Ye 1991), the two concepts being central to academic rhetoric. It also can pose a serious problem to EAP learners who "commonly have serious difficulties with the range of choices involved in reporting" (Thomas & Hawes 1994: 131). Therefore, studies on the use of reporting structures in research writing not only contribute to a better understanding of how individual disciplines and cultures construct argument, but are also valuable from a pedagogical point of view.

This work in progress report looks at the use of reporting verbs to express evaluative author stance from both cross-disciplinary and cross-linguistic perspectives. Based on a self-compiled comparable corpus (following guidelines by Connor & Moreno 2005; Moreno 2008) of 60 linguistic and literary research articles in English and Lithuanian, it aims to answer the question whether it is the "disciplinary culture" (Mauranen 1993) or the cultural tradition that prevails in the writers' choices of particular reporting verbs and the evaluative aspects they entail. The results obtained from the analysis of research articles are also compared to reporting verb choices made in 30 BA and MA papers on literature and linguistics written in English by Lithuanian students. The preliminary results point towards disciplinary rather than cultural trends in the choice of reporting structures. The study also confirms that effective reporting might be quite challenging for the learners of English even at an advanced academic level.

References

Connor, U. & A.I. Moreno. 2005. Tertium comparationis: a vital component in contrastive rhetoric research. In P. Bruthiaux, D. Atkinson, W. G. Eggington, W. Grabe & V. Ramanathan eds.. *Directions in Applied Linguistics*. Clevedon: Multilingual Matters, 153-64

Hyland, K. 2002. Activity and evaluation: reporting practices in academic writing. In J. Flowerdew ed. *Academic Discourse.* London: Longman, 115-30.

Hyland, K. 2011. Academic discourse. In K. Hyland & B. Paltridge eds. *The Continuum Companion to Discourse Analysis*. London: Continuum, 171-84.

Mauranen, A. 1993. Cultural differences in academic discourse – problems of a linguistic and cultural minority. In L. Löfman, L. Kurki-Suonio, S. Pellinen & J. Lehtonen eds. *The Competent Intercultural Communicator: AFinLA Yearbook*. Helsinki: AfinLA, 157-74.

Moreno, A.I. 2008. The importance of comparable corpora in cross-cultural studies. In U. Connor, E. Nagelhout & W. V. Rozycki eds. *Contrastive Rhetoric*. Amsterdam: Benjamins, 25-41.

Thomas, S. & T.P. Hawes. 1994. Reporting verbs in medical journal articles. *English for Specific Purposes,*13(2): 129-48.

Thomson, G. & Ye, Y. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics,*12: 365-82.

Reportive evidentials in English and Lithuanian: What kind of correspondence?

Aurelija Usonienė & Audronė Šolienė (Vilnius University)

The paper is concerned with lexical realizations of reportive evidentiality (Boye and Harder 2009; Boye 2012; Celle 2009; Wiemer 2007; Wiemer and Stathi 2010) across different discourse types and languages. An attempt is made to see how language specific the realizations and conceptualization of indirect evidentiality are by contrasting the findings of the analysis of the data collected from various monolingual and parallel corpora. One of the purposes of this contrastive analysis is to find out what kind of correspondence one can expect when dealing with the reportive sub-domain of the linguistic category of evidentiality.

The analysis is focused on the hearsay adverbials in English (*reportedly*, *allegedly*, *supposedly*) and Lithuanian (*neva, tariamai, esą*) as well as their bi-directionally established translation correspondences (comment clauses, complement-taking predicates, *as*-parentheticals, etc.), which can be illustrated in the following examples from Glosbe:

(1) EN-orig: *This is rather a conservative approach, although the Commissioner is* ***reportedly*** *a liberal - but let us get down to business*.
LT-trans: *Tai greičiausiai konservatoriškas požiūris, nors Komisaras,* ***kaip žinome****, yra liberalas, tačiau grįžkime prie reikalo*. **'as we know'**

(2) EN-orig: *Regarding the Community market's* ***alleged dependence*** *on external suppliers, it is considered that…*
LT-trans: *Apie tai, kad Bendrijos rinka* ***neva priklauso*** *nuo išorės tiekėjų, manoma, kad…* '…Community market **allegedly depends** upon**…'**

The present study is corpus-based and makes use of quantitative and qualitative methods of research (Aijmer 2002; Dyvik 2004; Johansson 2007; Hasselgård, Ebeling and Oksefjell Ebeling 2013; among others). The Lithuanian data have been drawn from the Corpus of the Contemporary Lithuanian Language (CCLL), namely from the news and fiction sub-corpora, and the Corpus of Academic Lithuanian (CorALit). The English data have been extracted from the British National Corpus (BYU-BNC). To establish translation correspondences between the items under study, a parallel bidirectional fiction corpus ParaCorpEN-LT (see Usonienė and Šolienė 2010), and a collection of translations from English into Lithuanian of EU documents – Glosbe (https://lt.glosbe.com/lt/en/) have been used. Cross-linguistic parallels were also tested in a self-compiled translation corpus Pilot-Multi-Para-Corp which is comprised of the translations of Lithuanian History into five languages (English, Polish, Russian, Spanish, and French).

Our preliminary findings indicate that both sets of hearsay adverbials have nearly the same frequency distribution across different discourse types in the two languages, however there is very weak equivalence in their translation correspondences.

References

Aijmer, Karin. 2002. Modal adverbs of certainty and uncertainty in an English-Swedish perspective. *Information Structure in a Cross-linguistic Perspective.* In H. Hasselgård, S. Johansson, B. Behrens & C. Fabricius-Hansen eds. Amsterdam: Rodopi, 97-112.

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

Boye, Kasper. 2012. *Epistemic Meaning. A Crosslinguistic and Functional-cognitive Study*. Berlin: Mouton de Gruyter.

Celle, A. 2009. Hearsay adverbs and modality. In R. Salkie, P. Busuttil & J. van der Auwera eds. *Modality in English – Theory and Description.* Berlin: Mouton de Gruyter, 269-94.

Dyvik, He. 2004. Translations as semantic mirrors: from parallel corpus to wordnet. In K. Aijmer & B. Altenberg eds. *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23).* Amsterdam: Rodopi, 311-26.

Hasselgård, H., J. Ebeling & S. Oksefjell Ebeling eds. 2013. *Corpus Perspectives on Patterns of Lexis*. Amsterdam: Benjamins.

Johansson, S. 2007. Seeing through multilingual corpora. In R. Facchinetti ed. *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi, 51–72.

Usonienė, A. & A. Šolienė. 2010. Choice of strategies in realizations of epistemic possibility in English and Lithuanian. *International Journal of Corpus Linguistics*, 15(2): 291-316.

Wiemer, B. & K. Stathi. 2010. The database of evidential markers in European languages. A bird's eye view of the conception of the database (the template and problems hidden beneath it). *STUF* 63(4): 275–289. Akademie Verlag.

Wiemer, B. 2007. Lexical markers of evidentiality in Lithuanian. *Rivista di Linguistica* 19(1): 173-208.

Data sources

BYU-BNC. Davies, M. (2004-) *BYU-BNC*. (Based on the British National Corpus, Oxford University Press). http://corpus.byu.edu/bnc/.

CCLL. Corpus of the Contemporary Lithuanian Language. http://tekstynas.vdu.lt/.

CorALit. Corpus of Academic Lithuanian. http://www.coralit.lt/.

Glosbe. Multilingual Online Dictionary (translation memory online: 1,013,284,995 translated sentences). https://glosbe.com/.

ParaCorp $_{EN-LT}$ Parallel Bidirectional English-Lithuanian Fiction Corpus.

Basic verbal communication verbs in Swedish and English from a crosslinguistic perspective

Åke Viberg (Uppsala University)

Speech act verbs or Verbal communication verbs (VCVs) as they will be referred to in this article represent one of the most extensive semantic fields of verbs in English and Swedish. The more or less complete inventory of VCVs has been thoroughly studied in English (Ballmer & Brennenstul 1981, Wierzbicka 1987; cf. for German Harras et al 2004, 2007). The FrameNet database contains representations of English verbs based on selected corpus examples (the Communication frame). A taxonomy of Swedish communication verbs has been proposed in Allwood (1977; cf. Allwood 1976, ch. 14). From a contrastive perspective, Proost (2007) looks at lexical gaps in the inventories of VCVs in English, German and Dutch.

There are also two studies that similar to the present one are based on translation corpora: Rojo & Valenzuela (2001) of Spanish to English translations and Shi (2008) of narratives in English and Chinese.

English and Swedish VCVs will be compared primarily on the basis of data from the English Swedish Parallel Corpus (ESPC) prepared by Altenberg & Aijmer (2000), which contains original texts in English and Swedish and their translations. Some data will also be taken from the Multilingual Parallel Corpus (MPC), which at present consists of extracts from 22 Swedish novels and their translations into English, German, French and Finnish (around 600 000 words in the Swedish originals).

The major focus of the present study is the most frequent verbs. In spite of the fact that there are around 400 Verbal communication verbs in the Swedish SUC-corpus (1 million words, mixed written genres), the most frequent verb *säga* 'say' accounts for 22% and the 10 most frequent verbs belonging to the field account for close to 50% of the textual occurrences of verbal communication verbs in this corpus. The most frequent verbs are also the most varied with respect to the range of constructions they can appear in and the patterns of polysemy that characterize them. Three studies of the most frequent English VCVs will be taken as a point of departure for the analysis. Dirven et al. (1982) introduced the concept of the linguistic action scene expressed in English with the four frequent verbs *say, tell, talk* and *speak*. There are four Swedish verbs that are frequent as translations, namely *säga, berätta, tala* and *prata*, but the functions of these verbs are distributed in a different way. In addition, tala 'speak' provides a good illustration of the exploitation of derivation and compounding in Swedish. Rudzka- Ostyn's (1989) study of ask accounts for a number of different meanings which are primarily distributed between two verbs in Swedish: *fråga* 'ask a question' and be 'request (politely)'. The third study taken as a point of departure (Rudzka-Ostyn 1995) is concerned with the network of meanings of the verbs of answering in English. In Swedish, the corresponding network is built around the verb *svara* and its uses as a derived and prepositional verb.

References

Altenberg, B. & K. Aijmer. 2000. The English-Swedish Parallel Corpus: a resource for contrastive research and translation studies. In C. Mair & M. Hundt eds. *Corpus Linguistics and Linguistic Theory*. Amsterdam. Rodopi, 15-23.

Allwood, J. 1976. *Linguistic Communication as Action and Cooperation.* Gothenburg Monographs in Linguistics 2, University of Göteborg, Dept of Linguistics.

Allwood, J. 1977. Om analys av kommunikationsverb. *Nysvenska Studier*, 57.

Ballmer, T. T. & W. Brennenstuhl. 1981. *Speech Act Classification: A Study in the Lexical Analysis of English Speech Activity Verbs*. Berlin. Springer.

Dirven, R., Goosens, L., Y. Putseys & E Voralt. 1982. *The Scene of Linguistic Action and its Perspectivization by Speak, Talk, Say and Tell*. Amsterdam: Benjamins.

Harras, G. et al. 2004. *Handbuch deutscher Kommunikationsverben. 1. Wörterbuch*. Berlin: de Gruyter.

Harras, G. et al. 2007 *Handbuch deutscher Kommunikationsverben 2. Lexikalische Strukturen*. Berlin: de Gruyter.

Proost, K. 2007. *Conceptual Structure in Lexical Items*. Amsterdam: Benjamins.

Rojo, A. & J. Valenzuela. 2001. How to say things with words: ways of saying in English and Spanish. *Meta* 44 (3): 467–77.
Rudzka-Ostyn, B. 1989. Prototypes, schemas and cross-category correspondences: the case of *ask*. *Linguistics*, 27, 613-61.
Rudzka-Ostyn, B. 1995. Metaphor, schema, invariance: the case of verbs of answering. In L. Goossens, P. Pauwels, B. Rudzka-Ostyn, A-M. Simon-Vandenbergen & J. Vanparys. 1995. *By Word of Mouth.* Amsterdam: Benjamins, 205-43.
Shi, D. 2008. Communication verbs in Chinese and English. a contrastive analysis. *Languages in Contrast*, 8(2): 181-207.
SUC 1.0. 1997 The Stockholm Umeå Corpus. Dept. of Linguistics, Umeå University and Dept. of Linguistics, Stockholm University. CD ROM.
Wierzbicka, A. 1987. *English Speech Act Verbs. A Semantic Dictionary*. Sydney: Academic Press.

Electronic source

FrameNet: http://www.icsi.berkeley.edu/~framenet/

Workshop IV: The future of the International Corpus of English (ICE): New challenges, new developments

Convenors: Robert Fuchs (University of Münster), Ulrike Gut (University of Münster) & Gerald Nelson (Chinese University of Hong Kong)

Speech and writing in South Asian Englishes: Corpus-based pilot studies on medium-related unity and diversity in Indian and Sri Lankan English

Tobias Bernaisch, Dilini Algama & Joybrato Mukherjee (University of Gießen)

While corpus-linguistic investigations of native varieties of English show convergence in writing and divergence in speech (cf. Mair 2007: 84), there is no consensus on the relation between speech and writing in South Asian Englishes. Their acquisitional parameters, which are more strongly characterised by formal, school-based instruction of the (written) language in contrast to the informal patterns of language acquisition at home in first-language varieties (cf. Kachru 1983: 41f.), have been argued to trigger a "register shift" (Mesthrie & Bhatt 2008: 114), i.e. the adoption of written norms in spoken discourse. According to Shastri (cf. 1988: 18), this may account for the bookish nature of South Asian Englishes. Still, Meyler (cf. 2007: xiv) argues that, based on the marked differences between written and spoken Sinhala and Tamil in Sri Lanka, the structural distance between written and spoken variants is higher in Sri Lankan English than in British English.

The present paper studies the relation between written and spoken variants of Indian and Sri Lankan English. The corpus data stem from the Indian component of the International Corpus of English (ICE-India), from the written part of ICE-Sri Lanka (cf. Körtvelyessy et al. 2012) and from a pilot version of the spoken component of ICE-Sri Lanka. Via three structural objects of investigation, i.e. postadjectival particle use (cf. Gunesekera 2005: 131; e.g. *different* + *from*/*to*/*than*), reduplications for intensification (e.g. "hot hot hoppers" (Meyler 2007: 116)) and invariant tag questions (cf. Lange 2012), we examine to what extent speech and writing in South Asian Englishes are characterised by a high degree of homogeneity (cf. Mesthrie & Bhatt 2008: 114; Shastri 1988: 18) or a high degree of diversity (cf. Meyler 2007: xiv).

For postadjectival particle use, the results highlight crossvarietal and cross-medium uniformity. In written and spoken Indian and Sri Lankan English, *different* is by default complemented with the particle *from* as in (1), although *different to* as in (2) figures more prominently as a minority variant in Sri Lankan than in Indian English.

(1) <w>it's</w> not going to be too <u>different from</u> your average day <ICE-SL:S2B-004#6:1:A>

(2) This is <u>different to</u> the British concept <ICE-SL:S1B-060#85:1:A>

Our findings show that reduplications and invariant tag questions occur more frequently in the spoken variants of both South Asian varieties than in their written variants, which highlights a clear distinction that South Asian speakers of both speech communities make between the two media. Our results entail implications for corpus compilation and future research. First, the variety-specific structures need to be considered in the compilation and annotation of an ICE component so that e.g. reduplications are not incorrectly marked as repetitions and the different invariant indigenous tag questions (e.g. *no* and *na* in Indian English (cf. Lange 2012: 204ff.)) are distinctively represented. Second, we stress the need to empirically study structural features in spoken and written variants of postcolonial Englishes – particularly those with acquisitional parameters comparable to South Asian Englishes – prior to making generalisations about the relation between speech and writing in (regional groupings of) postcolonial Englishes.

References

Gunesekera, M. 2005. *The Postcolonial Identity of Sri Lankan English*. Colombo: Katha Publishers.

Kachru, B. B. 1983. *The Indianization of English*. New Delhi: Oxford University Press.

Körtvelyessy, M., T. Bernaisch, J. Mukherjee & D. Mendis. 2012. *Manual to the Written Component of the International Corpus of English. Sri Lanka ICE-SL [W200]*. Giessen: Justus Liebig University.

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

Mair, C. 2007. British English/American English grammar: convergence in writing: divergence in speech? *Anglia*, 125, 84-100.

Mesthrie, R. & R. M. Bhatt. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.

Meyler, M. 2007. *A Dictionary of Sri Lankan English*. Colombo: Mirisgala.

Shastri, S.V. 1988. The Kolhapur Corpus of Indian English and work done on its basis so far, *ICAME Journal* 12, 15-26.

Expanding ICE to the expanding circle: The corpus of Dutch English

Alison Edwards (University of Cambridge)

This talk aims to make a threefold contribution: conceptual, methodological and empirical. First, it demonstrates that the scope of the International Corpus of English (ICE; Greenbaum, 1991) can – indeed, should – be widened to the Expanding Circle. ICE expressly includes only 'countries where [English] is either a majority first language … or an official additional language' (Greenbaum, 1996: 3); i.e. Inner and Outer Circle countries. This reflects a now outdated conception of English in the world based on its spread by way of colonisation. Today, the forces of globalisation mean that many people in Expanding Circle countries are using English comfortably and confidently well beyond the confines of the foreign language classroom. New corpora ought to reflect this development.

This talk describes the compilation of the Corpus of Dutch English (Edwards, 2011, 2014a), based on the design of the ICE corpora. For practical reasons it is presently limited to a written component, with 200 texts in 8 genres, totalling approximately 400,000 words. I

discuss the specific challenges involved in collecting text types such as English fiction, press news and social correspondence in the Netherlands, and the modifications to the ICE design required in such a setting. Further, I demonstrate how the Java-based platform Eclipse (http://www.eclipse.org) can be used to encode the corpus in XML and add metadata and textual markup in line with ICE (Nelson, 2002).

The talk then sums up the results of studies conducted with this corpus to date. The first is a study of the progressive aspect in the Corpus of Dutch English compared to the written components of four ICE corpora (Edwards, 2014b). No strict divide was found between the results for ICE-IND and ICE-SIN on the one hand and Dutch English on the other, suggesting that Outer and Expanding Circle varietal types should not be regarded as fundamentally different but as being on a continuum (see also Biewer, 2011: 28; Bongartz & Buschfeld, 2011: 48; Buschfeld, 2011: 219; Gilquin & Granger, 2011: 76).

This is supported by a second study (Edwards & Laporte, 2015) comparing preposition usage in Dutch English with five ICE corpora and four components of the International Corpus of Learner English (ICLE, Granger, 2003), including the ICLE component for the Netherlands (ICLE-NL). The Corpus of Dutch English and the ICE corpora clustered together, while, separately, the ICLE corpora (including ICLE-NL) clustered together. This suggests that (at least in terms of linguistic form), there is a false equivalence between 'Expanding Circle variety' and 'learner variety', and that users and learners can in fact co-exist in the Expanding Circle. It remains to be seen whether the Netherlands should be considered a special case, or whether it will be feasible to create comparable ICE-like corpora for other Expanding Circle countries to further test and extend these findings.

References

Biewer, C. 2011. Modal auxiliaries in second language varieties of English: a learner's perspective. In J. Mukherjee & M. Hundt eds. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: Benjamins, 7–33.

Bongartz, C., & S. Buschfeld. 2011. English in Cyprus: Second language variety or learner English? In J. Mukherjee & M. Hundt Eds., *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: Benjamins, 35–54.

Buschfeld, S. 2011. *The English Language in Cyprus: An Empirical Investigation of Variety Status*. PhD dissertation, University of Cologne.

Edwards, A. 2011. Introducing the Corpus of Dutch English. *English Today*, 27(03), 10-14.

Edwards, A. 2014a. *English in the Netherlands: Functions, Forms and Attitudes*. PhD dissertation, University of Cambridge.

Edwards, A. 2014b. The progressive aspect in the Netherlands and the ESL/EFL continuum. *World Englishes*, 33(2), 173-94.

Edwards, A., & S. Laporte. in press. Outer and Expanding Circle Englishes: the competing roles of norm orientation and proficiency levels. *English World-Wide*, 36.

Gilquin, G., & S. Granger. 2011. From EFL to ESL: evidence from the International Corpus of Learner English. In J. Mukherjee & M. Hundt eds. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: Benjamins, 55-78.

Granger, S. 2003. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), 538-46.

Greenbaum, S. 1991. ICE: The International Corpus of English. *English Today*, 7(4), 3-7.

Greenbaum, S. 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon.
Nelson, G. 2002. Markup Manual for Wwritten Texts. International Corpus of English. http://ice-corpora.net/ice/manuals.htm.

Exploring speaker fluency with phonologically annotated ICE corpora

Robert Fuchs & Ulrike Gut (University of Münster)

This presentation demonstrates how characteristics of speaker fluency can be measured in the phonologically annotated and time-aligned corpora ICE-Nigeria and ICE-Scotland. The former was compiled at the universities of Augsburg and Münster from 2009 to 2014; the latter is currently being constructed at the University of Münster. Both corpora belong to the 'new generation' of ICE corpora that include time-aligned transcriptions of the spoken part (Wunder et al. 2010). Both ICE corpora were/are compiled with the computer program *Pacx* (Platform for Annotated Corpora in XML; http://www.pacx.sf.net), a tool that combines linguistic data management, annotation, searching and distribution (Gut 2011). It extends the Eclipse platform (http://www.eclipse.org) by a set of tools, the XML editor Vex, the image viewer QuickImage and Subversive, which is a client for the version control system Subversion that supports collaborative work, and makes use of the software ELAN (http://www.lat-mpi.eu/tools/elan) for the annotation of audio and video files. Furthermore, some parts of the corpora were transcribed phonologically using the phonemic forced alignment system MAUS (Munich Automatic Segmentation System, Schiel 2004) for the automatic creation of time-aligned phonological transcriptions.

We show how these phonological transcriptions can be used for exploring speaker fluency in the ICE corpora. 10,000 words each of the categories broadcast talk and parliamentary debates in ICE Nigeria and ICE Scotland were analysed, and for each speaker the mean length of run (= average number of words per utterance) and articulation rate (=mean number of phonemes per utterance) were calculated. First results suggest some differences between the L1 Scottish and the L2 Nigerian speakers of English. Mixed effects regression models with SPEAKER as a random factor show that utterances in broadcast talks are significantly longer in Scottish than in Nigerian English ($p<0.01$), with a mean length of run of 7.1 words for Scottish and 5.7 words for Nigerian speakers. Articulation rate depends in both varieties mainly on the length of utterances, with longer utterances being spoken faster. When this is taken into account, Scottish speakers articulate still somewhat faster than Nigerian speakers (close to significance at $p=0.058$). For example, utterances that are five or six words long have a mean articulation rate of 11.4 phonemes/second in Nigerian and 12.3 phonemes/second in Scottish English. These results show how phonological and time-aligned annotation can enrich ICE corpora, and how they allow comparisons of varieties of English that go beyond studies of syntactic and lexical variation.

References

Gut, U. 2011. Language documentation and archiving with Pacx, an XML-based tool for corpus creation and management. In D. Nathan ed. *Workshop on Language Documentation and Archiving, London*, 21-25.

Schiel, F. 2004. MAUS goes iterative. *Proceedings of the IV. International Conference on Language Resources and Evaluation, Lisbon, Portugal*, 1015-18.

Wunder, E.-M., H. Voormann & U. Gut. 2010. The ICE Nigeria corpus project: creating an open, rich and accurate corpus. *ICAME Journal* 34, 78-88.

Using the ICE metadata for studying changes in the New Englishes: Is *must* decreasing in Hong Kong English?

Beke Hansen (University of Kiel)

The ICE metadata have only rarely been used to study intra-varietal variation in the New Englishes. Fuchs & Gut (2012), Höhn (2012), Lange (2012), Schweinberger (2012) and Fuchs & Gut (2015) are notable exceptions in this area. The neglect of the ICE metadata is probably conditioned by the rather time-consuming reprocessing of the data and metadata (esp. normalisation). However, considering the findings of the present study, the benefits of using the metadata outweigh the costs of reprocessing the data.

This paper illustrates the potential of the ICE metadata for studying changes in the New Englishes by investigating the use of the modal and semi-modal verbs of obligation and necessity in apparent time with data drawn from the ICE-HK metadata. Diachronic corpus-based research on BrE shows that the core modal *must* is decreasing, while the semi-modal *HAVE to* is increasing (Leech et al. 2009). My findings reveal that these diachronic trends are reflected in apparent time in ICE-GB. As there are no diachronic corpora for Hong Kong English (yet) and the ICE corpora are constructed to be comparable, it seems valid to use the same method for ICE-HK to test the widespread assumption that the New Englishes follow the trends of their input varieties (as proposed by Collins 2009, Collins & Yao 2012).

My findings show consistent variation in the use of the modal and semi-modal verbs of obligation and necessity according to age in the spoken component of ICE-HK. The core modal *must* is gradually decreasing from the oldest to the youngest age group, while the semi-modal HAVE *to* is on the increase in ICE-HK; in fact, the youngest age group strongly prefers semi-modal HAVE *to* over core modal *must*. The data also reveal gender-related variation, as women use the innovative semi-modal HAVE *to* more often than men.

The ICE metadata can furthermore be used to investigate claims about exonormative prescriptive pressure as an important factor in the early development of ESL varieties. The avoidance of informal variants is often ascribed to prescriptivism, and my findings substantiate this claim. In ICE-HK, (HAVE) *got to* is avoided by younger speakers, who probably adhere most closely to prescriptive norms which are taught at school. The findings also indicate that the informal variant is used more often by male speakers than female speakers – in contrast to HAVE *to*. The investigation of the semi-modal (HAVE) *got to* shows

another benefit of using the metadata, namely the possibility of uncovering idiosyncrasies in the data with the help of additional biographical speaker information.

In this way, my study reflects the "growing awareness that no variety is a monolithic entity and that intravarietal variation exists in all New Englishes" (Mukherjee & Schilk 2012:194). It complements large-scale inter-varietal studies by acknowledging the 'polylithic' nature of one variety, and shifts the focus to the variety itself by explaining variation from within rather than against the yardstick of the 'parent' variety.

References

Collins, P. 2009. Modals and quasi-modals in world Englishes. *World Englishes* 28/3: 281-92.

Collins, P. & X. Yao. 2012. Modals and quasi-modals in New Englishes. In M. Hundt & U. Gut eds. 2012. *Mapping Unity and Diversity World-Wide: Corpus-Based Studies of New Englishes*. Amsterdam: Benjamins, 35-54.

Fuchs, R. & U. Gut. 2012. Do women use more intensifiers than men? Investigating gender- and age-specific language use with the International Corpus of English. Paper presented at ICAME 33, Leuven, Belgium.

Fuchs, R. & U. Gut. 2015. An apparent time study of the progressive in Nigerian English. In P. Collins ed. 2015. *Grammatical Change in English World-Wide*. Amsterdam: Benjamins, 373-88.

Höhn, N. 2012. "And they were all like 'What's going on?'": New quotatives in Jamaican and Irish English. In M. Hundt & U. Gut eds. 2012. *Mapping Unity and Diversity World-Wide: Corpus-Based Studies of New Englishes*. Amsterdam: Benjamins, 263-90.

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

Leech, G., M. Hundt, C. Mair & N. Smith. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Mukherjee, J. & M. Schilk. 2012. Exploring variation and change in New Englishes: Looking into the International Corpus of English (ICE) and beyond. In T. Nevalainen & E .C. Traugott eds. 2012. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 189-99.

Schweinberger, M. 2012. The discourse marker LIKE in Irish English. In B. Migge & M. Ní Chiosáin eds. 2012. *New Perspectives on Irish English*. Amsterdam: Benjamins, 179-201.

ICE-corpora: Whence and whither? Alignment and annotation

John M. Kirk (Technische Universität Dresden)

Greenbaum's vision for creating a large international corpus enabling systematic comparisons across national varieties of English is largely being fulfilled. Once content (choice and quantity of text types) and methodology for compilation and markup had been agreed, storage and manipulation depended on mainframe computers. Beyond orthographic transcriptions, with only very general guidelines, which constituted the spoken component of the corpus, mark-up was ad hoc. Annotation amounted to part-of-speech tagging and syntactic parsing. Early critical discussion was concerned with the Britishness of the text categories and the difficulty of collection in L2 countries; countries for inclusion; questions of sampling and social representativeness; and funding.

In 2015, the corpus makes available some 14 (from over 20) national components, some dating from the original collection period of the early 1990s, others quite recent, with others still being compiled. Hardware and corpus-exploitation software have been revolutionized, greatly facilitating accessibility and convenience for researchers. In furtherance of Greenbaum's vision, this paper will be concerned with **audio & video alignment** and **pragmatic & prosodic annotation**. It will discuss the paradox which each development resolves; and address new questions now answerable.

It has been customary in corpus linguistics to accept transcriptions to be tantamount to the spoken language. Yet transcriptions are not spoken language – merely a transfer of an utterance from an audio-visual medium to an electronic written medium, with much of what is heard left out, and with many limitations. No-one has conceived of studying speech for descriptive purposes without using a transcription, and no-one has suggested search and display mechanisms purely in terms of audio or visual signals, as if it were possible to construct a strictly audio or strictly visual concordance. The paradox is reconcilable through **alignment**. Audio- and visual recordings have been aligned to transcriptions, as with ICE-GB, and it seems likely that future ICE corpora will be audio-aligned as well. Whereas transcriptions are based on the audio-recording, the alignment of the audio-recording depends on the transcription. Thus through the transcription it is possible to get back to the original recording and hear it replayed in tandem with the transcription. With audio- or video-alignment as well, it is now possible to get closer than ever to actual real speech.

The paradox is also reconcilable through further **annotation** beyond that for parts-of-speech and syntactic parsing. With prosodic and pragmatic annotation, as in the pioneering SPICE-Ireland Corpus, it is possible to objectify in the transcription the illocutionary intentions as well as some of the perlocutionary effects of an utterance for an altogether deeper understanding of how communicative exchanges work. Key features are the many discourse markers associated with illocutionary force and a speaker's stance towards their utterance and their audience. By encapsulating such information, some of it expressed by pitch movement, some by discourse markers, others by lexical and syntactic choices, the annotation complements and upgrades the transcription as a richer objectification of the original spoken utterance. What transcription has traditionally left out, audio-video alignment and multi-categorial annotation can thus effectively now put back in.

ICE Online

Hans Martin Lehmann (University of Zurich)

In our paper we present ICE Online, a framework for the search and analysis of the ICE Corpora (cf. http://www.es.uzh.ch/dbank). We discuss how different types of linguistic and contextual annotation can be used to define, classify, restrict and tabulate search results.

We have integrated ***linguistic annotation*** from two projects, dependency syntax (cf. Lehmann and Schneider 2012) and pragmatic annotation (cf. Kallen and Kirk 2012). We will

show how rich layers of linguistic annotation can be accessed in the interface. SPICE Ireland can be searched not only via part of speech, lemma, verb/noun chunks and syntactic annotation, but also by means of pragmatic annotation. We look at the structural/segmental information present in the distributed corpora and the additional segmentation imposed by the annotation process. We will also discuss how the pragmatic annotation provided in SPICE was integrated into our system.

In addition, the framework presented here also provides access to ***contextual annotation***. For searches across the ICE Corpora we will show distributions derived from the ICE sampling frame. For ICE GB and ICE Ireland we will make use of the contextual annotation provided by the corpus compilers, e.g. *sex, age, religion, region,* etc. The system we present here can automatically calculate word-frequencies for ad-hoc tabulations and cross-tabulations of contextual variables and is thus capable of providing relative frequencies for individual cells.

Our framework also offers preformatted databases for annotating search results manually with ***ad-hoc classifications***. This is achieved without losing access either to the distributed co-text or to the contextual and linguistic annotation. As a matter of course, our system depends on the corpus-compilers to provide contextual data and could be put to much better use if the contextual annotation were standardized across the ICE Corpora.

References

Lehmann, H. M. & G. Schneider. 2012. Dependency Bank. In V. B. Mitielu, O. Popescu & V. Pekar. eds. *LREC 2012 Challenges in the Management of Large Corpora*. Istanbul. 67-77.
Kallen, J. L & J. M. Kirk. 2012. *SPICE-Ireland: A User' Guide*. Belfast: Cló Ollscoil na Banríona.

ICE (International Corpus of English) vs GloWbE (Corpus of Global Web-based English): A critical approach to big data

Lucía Loureiro-Porto (University of the Balearic Islands)

The validity and usefulness of the ICE corpora is now, more than 20 years after the project was launched, completely unquestionable. Thousands of publications on New Englishes (McArthur 1992: 688), Postcolonial Englishes (Schneider 2007) or, as they are most frequently referred to nowadays, World Englishes (Mesthrie and Bhatt 2008) have appeared in the most influential journals and book series using different ICE corpora as sources of data. Nevertheless, its size, one million words, appears to be too limited for 21st century linguistics. The coming of age of big data seems to call for larger corpora which include millions of words (e.g. Bieman *et al*. 2013). Just like the sequencing of the genome and the development of subsequent omics (e.g. genomics, proteomics, etc.) revolutionized Biology, because it made trillions of data available to the researchers, the easy access to Internet material may become an analogous landmark for Linguistics. That appears to be the case with Mark Davies' *Corpus of Global Web-based English* (GloWbE), which contains 1.9 billion words from 20 different countries. As Davies and Fuchs (2015) show, this corpus is valid for the study of variation at the lexical, morphological, syntactic, semantic and even pragmatic

level. Although the same can be said about ICE, there is no doubt that each of these corpora represents one of the two classes proposed by Rissanen (2000: 8): ICE qualifies as small-sized "multi-purpose" corpus, while GloWbE (compiled mainly with webpages and blogs) falls within the class of larger but genre-specific corpora. Size, however, is a more than debated feature of corpora, and the larger does not necessarily imply the better (as shown, for example, by Nurmi 2002 for two historical corpora). In this scenario, the aim of this paper will be to establish a comparison between ICE and GloWbE for four varieties of English, namely British English (for which the *British National Corpus* will be used as a gold standard) and the Englishes in India, Hong Kong and Singapore. The linguistic variables that will be considered are: (a) frequency and collocations of modal verbs, (b) presence of colloquial features (e.g. hedges, fillers, contractions, etc.), and (c) relative frequency of local vocabulary items. The differences between both corpora regarding each of the linguistic variables will be quantitatively measured using basic statistical tests, such as Kennedy's (1998) coefficient of difference, log-likelihood functions and Pearson's correlation coefficient, among others. The results will show that, despite its much larger size, GloWbE is not necessarily the best corpus to choose when conducting a study on linguistic variation in World Englishes. The much smaller ICE corpora may constitute more representative samples of these varieties of English. Recovering the comparison with the omics, it is fitting to say that Biology is already witnessing the rise of skeptical voices as for the use of large amounts of data and calling for more detailed methodologies (e.g. Hanage 2014). Perhaps Linguistics should also reflect on the most suitable role of Internet material in variationist studies, rather than falling into the temptation of using it as the main source for corpora.

References

Biemann, C., F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, J. Simon, L. Swiezinski & T. Zesch. 2013. Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics* 28(2): 23-59.

Davies, M. 2013. *GloWbE (Corpus of Global Web-Based English)*. http://corpus2.byu.edu/glowbe/. Accessed 10 Feb 2015.

Davies, M. & R. Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 Billion Word Global Web-Based English Corpus (GloWbE). *English World-Wide*, 36(1). [In Press.]

Hanage, W. 2014. Microbiome science needs a healthy dose of scepticism. *Nature* 512: 247-48.

ICE project (International Corpus of English). 1990-present. http://ice-corpora.net/ice/. Accessed 30 April 2013.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London & New York: Longman.

McArthur, T. 1992. *The Oxford Companion of the English Language*. Oxford: OUP.

Mesthrie, R. & R. M. Bhatt. 2008. *World Englishes. The Study of New Linguistic Varieties*. Cambridge: CUP.

Nurmi, A. Does size matter? The *Corpus of Early English Correspondence* and its sampler. In H. Raumolin-Brunberg, M. Nevala, A. Nurmi & M. Rissanen eds. *Variation Past and Present. VARIENG Studies on English for Terttu Nevalainen*. Helsinki: Société Néophilologique, 173-84.

Rissanen, M. 2000. The world of English historical corpora. From Cædmon to the Computer Age. *Journal of English Linguistics* 28(1): 7-20.

Schneider, E. W. 2007. *Postcolonial English. Varieties round the World*. Cambridge: CUP.

New insights into existing ICE data with a new corpus architecture

Simon Sauer (Humboldt-Universität zu Berlin)
John M. Kirk (Technische Universität Dresden)

Most of the ICE subcorpora available today are provided only in plain text files, marked up according to the standardized ICE guidelines. This format certainly has its benefits in providing a low-tech software-independent solution while at the same time guaranteeing compatibility to all previous ICE releases. However, technological advances have made it possible to parse and annotate large amounts of data (semi-)automatically in a vast variety of different categories. Moreover, digital audio recordings and ever-increasing bandwidths have made it feasible to publish (transcription-aligned) audio and even video data. On the other hand, the more information is put into a corpus, the less legible the original text or transcription becomes, both for human readers and for corpus analysis tools.

The recent ICE Nigeria has opted to make use of several formats, including ELAN files that are time-aligned with the audio recordings. This makes it much easier to look at the data and allows for detailed qualitative analyses. Quantitative analyses, on the other hand, are faced with a number of difficulties. Analyses across several ICE subcorpora require different approaches, as the ICE mark-up has not been applied. Annotations and the time-alignment present in the ELAN files cannot be used with standard corpus analysis tools.

This paper will present the benefits of a generic multilayer corpus architecture that allows multiple, independent annotations of any type (including time-alignment) as well as the inclusion of metadata. At the core of this architecture lies SaltNPepper (1), which consists of a generic meta-model and an extendable converter framework that allows data from a variety of formats and tools, e.g. ELAN or generic XML. The data can then be converted into the format of ANNIS (2), an open source, web browser-based search and visualization interface.

We have converted ICE Ireland (3), spoken and written, as well as the pragmatically and prosodically annotated SPICE Ireland into this format and can therefore present the benefits of this approach by directly comparing it to the standard plain text format (hitherto the only format available for ICE Ireland).

ANNIS provides powerful query capabilities across different (meta)annotation types and even across corpora. This not only allows for complex queries but also for the creation of ad-hoc subcorpora using the metadata, so different groups of documents or speakers can easily be compared or results can be restricted to a specific subset. So far, the only way to filter results was to manually look up every single relevant speaker ID in the corpus handbook.

Query hits can be displayed in a variety of visualizations, including e.g. a grid-style view that displays annotations in separate layers, colour-coded text or syntax trees, as well as downloaded for further (statistical) analysis. Aligned audio/video data can be played back by clicking on any token or annotation. Even spoken subcorpora that can only provide transcriptions benefit greatly by being able to indicate overlaps visually while at the same time showing the transcription proper free from all the mark-up clutter.

References

(1) SaltNPepper: http://korpling.german.hu-berlin.de/saltnpepper.
F. Zipser & L. Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In G. Budin, L. Romary, T. Declerck & P. Wittenburg eds. *LREC 2010 Workshop, Proceedings, W4: Language Resource and Language Technology Standards*. Paris: ELRA. http://hal.inria.fr/inria-00527799.
(2) ANNIS: http://annis-tools.org
A. Zeldes, J. Ritz, A. Lüdeling & Ch. Chiarcos. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. In M. Mahlberg, V. González-Díaz & C. Smith eds. *Proceedings of Corpus Linguistics 2009*. http://edoc.hu-berlin.de/docviews/abstract.php?id=36996
(3) ICE- and SPICE-Ireland: http://ice-corpora.net/ice/iceire.htm.

Challenges of compiling a corpus of spoken English in Namibia

Helene Steigertahl (Universität Bayreuth)

As already noted by many linguists of World Englishes, it is not in all cases easy to classify national varieties of English according to the ENL-ESL-EFL distinction (Biewer 2011, Buschfeld 2013, Edwards 2014). Terms like "hybrid cases" and "non-PCEs" have been used but no clear label has yet been found. Investigating English spoken in Namibia faces the same problem. Undoubtedly, it is used as a NL, SL and FL in Namibia. Of course, age plays an important rule here, especially concerning Namibians who grew up before and after independence respectively. However, there are no statistics available on this. The term I came up with so far is "Independent English", referring to Namibia's historical implementation of English as the sole official language in 1990.

By the time of implementation, only 4% of the Namibian population were L2 speakers of English and only 0.8% were L1 speakers of English. Despite this fact, English was introduced as the only official language, in contrast to other African countries where English has served as one of several official languages, making the Namibian language policy exceptional.

Since its implementation the number of English L1 speakers has only slightly increased, up to 1.9% in 2001 and 3.4% in 2011. Nevertheless, English has spread through all private and public domains and is spoken by many people in the whole country as the attitudes towards this Indo-European language are very positive. In spite of that, English in Namibia, often called "Namlish" by the Namibians, has scarcely gained any attention by scholars of the New Englishes.

After 25 years in use, English in Namibia can be assumed to emerge as a new variety. However, it is yet unclear whether this English can be considered a sub-variety of South African English, a second language variety or rather a learner English. It seems as if Namibian English does not yet fit into any category of varieties. This again is a challenge for compiling a corpus of English in Namibia.

After giving a short overview of the linguistic situation of the country, I will present my methodological framework for compiling a corpus of spoken English in Namibia. This is my dissertation project and not part of the ICE Namibia initiative. Then, a number of findings from two field trips to Namibia in 2013 and 2015 will be depicted. The data was collected in the towns and surrounding areas of Windhoek, Usakos, Ruacana, Divundu, Gobabis and Luderitz. This already covers many areas of the country. Nonetheless, it seems almost impossible to conduct a large-scale corpus that includes speakers of every part of Namibia. This challenges the construction of a corpus once more.

This presentation will cover phonological and morpho-syntactic features of the "hybrid case" of English spoken in Namibia. Moreover, it will be shown that language use and attitudes are closely intertwined with the status and role of English in Namibia. Therefore, I suggest to consider attitude studies for corpus studies in the future. This could be a recommendation for the ICE project in order to address recent change in the global usage of English as NL, SL and FL.

References

Bauer, L. 2002. *An Introduction to International Varieties of English*. Edinburgh: Edinburgh UP.

Biewer, C. 2011. Modal auxiliaries in Second Language Varieties of English: a learner's perspective. In J. Mukherjee & M. Hundt eds. *Exploring Second-Language Varieties of English and Learner Englishes. Bridging a Paradigm Gap.* Amsterdam: Benjamins. 7-33.

Buschfeld, S. 2014. English in Cyprus and Namibia: a critical approach to taxonomies and models of World Englishes and second language acquisition research. In S. Buschfeld et al eds. *The Evolution of Englishes. The Dynamic Model and Beyond*. Amsterdam: Benjamins. 181-202.

Buschfeld, S. & A. Kautzsch. 2014. English in Namibia. A first approach. *English World-Wide*, 35*:2*. Amsterdam: Benjamins. 121-60.

Central Intelligence Agency (CIA). 2014. *The World Factbook*. *Africa: Namibia*. Washington, DC. https://www.cia.gov/library/publications/the-world-factbook/geos/wa.html. Accessed 19. Nov. 2014.

Chamberlain, R. 1981. *Toward a Language Policy for Namibia: English as the Official Language; Perspectives and Strategies*. [Namibia Studies Series; 4]. Lusaka: United Nations Institute for Namibia.

Edwards, A. 2010. Dutch English: tolerable, taboo, or about time too? On keeping versus 'correcting' Dutch Flavour in English texts. *English Today*, 26, 19-25.

Edwards, A. 2014. The progressive aspect in the Netherlands and the ESL/EFL continuum. *World Englishes*, *33*(2), 173-94.

Frydmann, J. 2011. A critical analysis of Namibia's English-Only language policy. In E. G. Bokamba et al 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. Accessed 8. Sept. 2014.

Harlech-Jones, B. 1995. Language policy and language planning in Namibia. In M. Pütz ed. *Discrimination through Language in Africa? Perspectives on the Namibian Experience*. Berlin/New York: Mouton de Gruyter. 181-206.

Harlech-Jones, B. 1997. Looking at means and ends in language policy in Namibia. In M. Pütz ed. *Language Choices: Conditions, Constraints, and Consequences.* Amsterdam: Benjamins. 223-49.

Lanham, L. 1996. A history of English in South Africa. In V. de Klerk ed. *Focus on South Africa*. Amsterdam: Benjamins. 19-34.

Lass, R. 2004. South African English. In R. Hickey ed. *Legacies of Colonial English. Studies in Transported Dialects*. Cambridge: CUP. 363-86.

Maho, J. F. 1998. *Few People, Many Tongues. The Languages of Namibia*. Windhoek: Gamsberg Macmillan.
McArthur, T. 1998. *The English Languages*. Cambridge: CUP.
Mollin, S. 2006. *Euro-English. Assessing Variety Status*. Tübingen: Gunter Narr.
Namibia Statistics Agency. 2001. *Namibia 2001. Population and Housing Census Main Report.* http://www.nsa.org.na/files/downloads/a5d_Namibia_2001_Population_and_Housing_Census_Main_Report.pdf. Accessed 19. Nov. 2014.
Namibia Statistics Agency. 2011. *Namibia 2011. Population and Housing Census Main Report.* http://www.nsa.org.na/files/downloads/Namibia_2011_Population_and_Housing_Census_Main_Report.pdf . Accessed 19. Nov. 2014.
Platt, J. T., H. Weber & M. L. Ho. 1984. *The New Englishes*. London, New York: Routledge & Kegan Paul.
Pütz, M. 1995. Attitudes and language: an empirical investigation into the status and use of English in Namibia. In M. Pütz ed. *Discrimination through Language in Africa? Perspectives on the Namibian Experience*. Berlin/New York: Mouton de Gruyter. 245-84.
Schneider, E. 2007. *Postcolonial English. Varieties around the World*. Cambridge: CUP.
Schneider, E. 2014. New reflections on the evolutionary dynamics of World Englishes. *World Englishes* 33(1), 9-32.
Strang, B. M. H. 1970. *A History of English*. London: Methuen.
Wallace, M. & J. Kinahan. 2011. *A History of Namibia. From the Beginning to 1990*. Johannesburg: Jacana Media.

Plenaries

"Magnificence of promises" in the Burney Collection

Kate Burridge (Monash University)

> WHATEVER is common is despised. Advertisements are now so numerous that they are very negligently perused, and it is therefore become necessary to gain attention by magnificence of promises, and by eloquence sometimes sublime and sometimes pathetic.
>
> PROMISE, large Promise, is the soul of an Advertisement. (Johnson 1761, 225; 1963, 125)

Georgian England saw the rapid and widespread growth of consumerism and the rise of commercial advertising as we know it — it was the golden age of puffery, of quackery and of humbuggery and Dr Samuel Johnson was a keen observer and critic. As he wrote in his essay *Art of Advertising* (first published in *Idler* 20th January 1759): "Promise, large Promise, is the soul of an Advertisement".

As the new medium of the day, newspapers were being filled by advertisements — "grown up by slow degrees to its present state" (to quote Johnson). In the early 1730s, the *London Daily Advertiser* featured fifteen advertisements, but by the 1780s carried around 300 ads per issue. As much as 75% of the space in some dailies in 1750 was devoted to advertisements; indeed, it was during this time that the word *Advertiser* began to replace *Post* in the names of publications (Turner 1965, 28). Newspapers allowed the flourishing of advertising, and advertising revenue propped up the newspapers — many would have gone bankrupt without it.

Curiously though, histories of advertising (such as Barrés-Bake 2006; Briggs 1993; Turner 1965; Walker 1973) dismiss this era as mere prehistory, claiming that we have to wait for the Victorians for "real" advertisements to arrive. To dismiss this claim, I will use evidence from the ads in the Burney Collection. This online database is comprised of the roughly 700 bound volumes of 17th and 18th century newspapers and news pamphlets (mostly published in London) originally collected by the Reverend Charles Burney (1757-1817). It is now the largest single collection of news media from this period, containing over 1,270 titles and around a million pages. In order to throw light on the language of eighteenth-century promotion, I will focus on the distinctive features of the noun phrase, in particular its adjectival hype or 'puff', as well as the metaphors that dominate this puff.

While at first blush these early advertisements appear to have little in common with what we find today (with their long-winded texts and dearth of graphics), to my mind the evidence is overwhelming that the promotional strategies and essential features of marketing language today were well and truly in place, curbed only by the limited media options available at the time — the linguistic trail left by the language of eighteenth-century promise-making leads directly to the "Gruen Transfer" of modern times.

References

Barrés-Baker, M. C. 2006. *An Introduction to the Early History of Newspaper Advertising*. Brent Museum and Archive Occasional Publications, no. 2. London: Brent Museum.

Briggs, Peter M. 1993. "'News from the Little World': A Critical Glance at Eighteenth-Century British Advertising". *Studies in Eighteenth-Century Culture* 23: 29–45.
Johnson, S. 1761. "The Idler. No. 40". In *The Idler*. London: Printed for J. Newbery, 224–29.
———. 1963. *The Idler and The Adventurer,* ed. W. J. Bate, John M. Bullitt and L. F. Powell. New Haven and London: Yale University Press.
17th–18th Century Burney Collection Newspapers.
Turner, E. S. 1965. *The Shocking History of Advertising*. Harmondsworth: Penguin.
Walker, R. B. 1973. "Advertising in London Newspapers, 1650–1750". *Business History* 15 (2): 112–30.

Corpora – constructions – cognition

Thomas Herbst (FAU Erlangen-Nürnberg)

Corpus linguistic research has resulted in a view of language that underscores the importance of multi-word units in language use, which led Sinclair to formulate the idiom principle. It is the aim of this paper to explore to what extent these findings can be accommodated in a construction grammar framework and to raise a few issues that are relevant in this context with a focus on phraseological and valency phenomena. Furthermore, it will be shown that such an approach may entail the need to throw some traditional categories overboard.

Contexts for contentful constructionalization

Graeme Trousdale (Edinburgh University)

Studies of grammaticalization phenomena have regularly emphasized the importance of context in the development of new grams (Bybee, Perkins and Pagliuca 1994); similarly, context has been seen as important factor in the accurate description and classification of constructional taxonomies (Bergs and Diewald 2009). In bringing together work on constructions and language change, research concerning the development of new procedural constructions (such as future tense constructions (Hilpert 2008), or cleft constructions (Traugott 2008, Patten 2012)) has also attended to the contexts in which the new constructions develop. However, the relationship between context and change in the creation of new contentful constructions (e.g. word-formation schemas) has been less fully explored. This talk provides an overview of work on the development of contentful constructions. It includes a discussion of both quantitative (e.g. Hilpert 2013, 2015) and qualitative (e.g. Traugott and Trousdale 2013, 2014) approaches to change in word-formation patterns, and considers how both cotext and context are relevant for understanding the development of new lexical schemas and micro-constructions. Drawing on data from a range of corpora, the development of different kinds of contentful constructions is considered, which is important because the nature of context may be different in the development of different kinds of constructions. For instance, what constitutes a relevant context in the development of a morphological construction (such as a particular kind of NN compounding, or the creation of nominalizing affixes in word-formation schemas) is different from that which constitutes the

context for a new or modified idiom or snowclone (such as *the proof is in the pudding* and *he's not the hottest biscuit in the oven*). In particular, the talk focuses on the importance of network links in the creation of new contentful constructions, and argues that in onset contexts for change, what is critical are tiny readjustments in the local constructional network, which is most strongly affected by spreading activation. The final part of the presentation considers this issue of spreading activation in more detail, in particular for the development of new word-formation schemas. It is suggested that spreading activation in networks (Hudson 2007) is as important in the development of contentful constructions as it is in procedural constructions, and that the link between spreading activation and priming is relevant in both types of constructional change. Hudson (2007: 40) suggests that processing is concerned with "finding the best 'path' from the (known) form to the (unknown) meaning", and it has been proposed that this path finding is enabled not only by priming, but also by the "implicatures and inferences made and accepted by participants in that discourse" (Traugott and Trousdale 2013: 55), highlighting the important of context. While this is particularly salient in the case of procedural constructionalization, it is shown also to be important in the development of new contentful constructions.

References

Bergs, Alexander & Gabriele Diewald. 2009. Introduction: contexts and constructions. In Alexander Bergs & Gabriele Diewald eds., *Contexts and Constructions*. Amsterdam: Benjamins, 1-14.

Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World*. Chicago: The University of Chicago Press.

Hilpert, Martin. 2008. *Germanic Future Constructions: A Usage-Based Approach to Language Change*. Amsterdam: Benjamins.

Hilpert, Martin. 2013. *Constructional Change in English: Developments in Allomorphy, Word-Formation and Syntax*. Cambridge: CUP.

Hilpert, Martin. 2015. From *hand-carved* to *computer-based*: noun-participle compounding and the upward-strengthening hypothesis. *Cognitive Linguistics* 26: 1-36.

Hudson, Richard. 2007. *Language Networks: the New Word Grammar*. Oxford: OUP.

Patten, Amanda L. 2012. *The English* it-*cleft: A Constructional Account and a Diachronic Investigation*. Berlin: De Gruyter Mouton.

Traugott, Elizabeth Closs. 2008. 'All that he endeavoured to prove was …': on the emergence of grammatical constructions in dialogic contexts. In Robin Cooper & Ruth Kempson, eds. *Language in Flux: Dialogue Coordination, Language Variation, Change and Evolution*. London: Kings College Publications, 143-77.

Traugott, Elizabeth Closs & Graeme Trousdale. 2013. *Constructionalization and Constructional Changes*. Oxford: OUP.

Traugott, Elizabeth Closs & Graeme Trousdale. 2014. Contentful constructionalization. *Journal of Historical Linguistics* 4: 256-83.

Accessing the Electronic Oxford English Dictionary

Edmund Weiner (Oxford University / OED)

When the digitization of the OED was undertaken in the mid 1980s, a transformation of the user's experience of the dictionary was envisaged. We expected that changes would occur in the way information that is unobtainable from the print dictionary could be extracted and in the way the content of the dictionary, accessed as the outcome of either simple or complex searches, was visualized and presented. There was also the possibility of extending the ways in which users can frame queries to the dictionary. This paper looks at the progress that has been made in these areas.

PETER LANG
BERN · BERLIN · BRUXELLES · FRANKFURT AM MAIN · NEW YORK · OXFORD · WIEN

Daphné Kerremans

A Web of New Words

A Corpus-Based Study of the Conventionalization Process of English Neologisms

Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien, 2015. 278 pp., 21 tables, 50 graphs
English Corpus Linguistics. Vol. 15
Edited by Thomas Konen and Joybrato Mukherjee
hb. ISBN 978-3-631-65578-8
CHF 68.– / €[D] 59.95 / €[A] 61.60 / € 56.– / £ 45.– / US-$ 72.95
eBook ISBN 978-3-653-04788-2
CHF 71.65 / €[D] 66.64 / €[A] 67.20 / € 56.– / £ 45.– / US-$ 72.95
€[D] includes VAT – valid for Germany and EU customers without VAT Reg No · €[A] includes VAT – valid for Austria

This book presents the first large-scale usage-based investigation of the conventionalization process of English neologisms in the online speech community. The study answers the longstanding question of how and why some neologisms become part of the English lexicon and others do not. It strings together findings and assumptions from lexicological, sociolinguistic and cognitive research and supplements the existing theories with novel data-driven insights. For this purpose a webcrawler was developed, which extracted the occurrences of the neologisms under consideration from the Internet in monthly intervals. The book shows that the different courses conventionalization processes may take result from the interplay between speaker-based sociopragmatic accommodation-induced aspects and factors facilitating cognitive processing of novel linguistic material.

CONTENTS: Neologisms • Lexical Innovation • Diffusion • Conventionalization • Cognition • Mental Lexicon • Online Speech Community • Sociopragmatic Analysis • Collocation • Emergence of Syntagmatic Networks • Accommodation • Webcrawler • Conventionalization Continuum • Nameworthiness.

Peter Lang AG · Internationaler Verlag der Wissenschaften
Moosstrasse 1 • Postfach 350 • CH-2542 Pieterlen • Schweiz
Tel: +41 (0)32 376 17 17 • Fax: +41 (0)32 376 17 27
order@peterlang.com **www.peterlang.com**

Full papers

Intensifiers across some national varieties of English

Karin Ajmer (University of Gothenburg)

Lately we have experienced a surge of interest in how language is used for evaluative functions and how evaluation is associated with particular linguistic elements. Intensifiers such as *so, very* and *really* are closely associated with the speaker's desire to use and exploit the expression of hyperbole or exaggeration for certain effects. Consider examples such as *I'm so fed up* where *so* +adjective expresses affect and is used to turn up the emotional volume in the on-going discourse.

The aim of my contribution is to investigate the distribution of the amplifying intensifiers *very*, *really* and *so* and their patterns of usage with adjectives in some national varieties of English. The focus is on the following research questions:

- What are the frequencies of the intensifiers studied in different varieties of English?
- What adjectives and how many adjectives do the intensifiers occur with?
- Can the patterns of usage of the intensifiers in different varieties related to the discussion of on-going changes, innovation and grammaticalization?

As has been shown in previous work the frequency of use and patterns of usage with intensifiers may differ between varieties of English (see eg Tagliamonte and Roberts 2005 on Canadian English; de Klerk 1995 on Xhosa English and New Zealand English, Coronel on Philippine English). The data in my study will consist of the spoken varieties of English in the ICE corpora from Great Britain, New Zealand and Singapore. In addition the Santa Barbara Corpus of American English (SBCAE) has been used in order to make comparisons between American English and the other varieties. It is shown that the frequency of *very, really* and *so* differs in the varieties.

Table 1. The ranking order of some frequent intensifiers in five regional corpora. Normalized frequencies (to 100,000 words) within parentheses

	SBCAE	ICE-GB	ICE-NZ	ICE-SIN
very	163 (65.7)	519 (288)	107 (59.4)	749 (416)
really	150 (60.5)	171 (95)	312 (173.3)	72 (40)
so	208 (83.9)	104 (57.8)	62 (34.4)	286 (158.9)

Singapore English uses both *very* and *so* more frequently than the other varieties. This overuse is however balanced by the 'underuse' of *really. So* was also frequent in American English. A tendency of focusing especially on two (amplifying) intensifiers (*really* and *very)* is characteristic of New Zealand English.

The adjectives collocating with the intensifiers can be described semantically along the parameter good-bad (although some adjectives are not evaluative). In Jim Martin's theory of appraisal (Martin 2000) a distinction is made between Affect, Judgement and Appreciation. It

is shown that there were differences between the varieties both with regard to the type of adjective and whether they collocated with a positive or negative term (Partington 2004). On the other hand, the most frequent adjectives were the same in the different varieties although they did not have the same frequencies. There were also differences in the range of adjectives used. Another difference was the extent to which the intensifiers were found with innovating and trendy adjectives in the different varieties.

References

Coronel, L. [Unpublished]. Patterns of intensifier usage in Philippine English.

De Klerk, V. 1995. Expressing levels of intensity in Xhosa English. *English World-Wide*, 26 (1): 77-95.

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.

Partington, A. 2004. "Utterly content in each other's company": semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9 (1): 131-56.

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

The LACELL project for collocation-based extraction of conceptual networks: Theory, methodology, results

Moisés Almela & Pascual Cantos (Universidad de Murcia)

This paper summarizes the theoretical underpinnings, the methodological decisions and the results of a three-year long research project in corpus-based lexical semantics carried out by members of the LACELL research group. The main goal of the project has been to develop a technique for describing the conceptual networks formed by patterns of lexical collocation in large-scale language corpora. The specific contribution of our approach, in comparison to the received descriptions of collocation in corpus linguistics, resides in having devised a system for detecting and describing patterns of dependency between different collocations of a word.

For instance, while a conventional approach to collocation in corpus linguistics would conclude that *dire* is a collocate of *consequence* if their co-occurrences in a corpus are statistically significant, our approach is not content with this conclusion and goes on to inquire about the possible influence exerted by other words frequently co-occurring with both *dire* and *consequence*. To capture the interaction between collocates we rely on conditional probabilities.

An important upshot of this line of inquiry is that the concept of collocation itself is in need of substantial revision. The mainstream definitions of collocation have failed to take into account the phenomenon of "multiple dependency," which is anything but marginal. In many cases the occurrence of a collocate in the context of the node is a factor of the attraction exerted not only by the node but also by other lexical items occurring in the same textual window. Thus, if there is evidence indicating that the attraction of *dire* towards the context of

consequence is determined not only by the noun itself but also by the verb of which *consequence* is direct object, the conclusion that *dire* is a collocate of *consequence* appears to be an oversimplification. Arguably, a more insightful conclusion is that in the context of *consequence*, *dire* is a "co-collocate" of specific verbs (e.g., *suffer, face, predict*).

The semantic relevancy of the distinction between co-collocates and collocates and its utility for the construction of conceptual networks are brought out at a higher-level of analysis, when the sets of individual lexical items found in specific slots of collocates and co-collocates are generalized to semantic types. Thus, to continue with the aforementioned example, those modifiers of *consequence* that convey a 'negative evaluation' (e.g., *dire, tragic, serious*) tend to co-collocate with verbs that contain the meaning 'be affected by' but not with verbs denoting an 'intellectual process' (e.g., *consider, explore, assess, evaluate, understand, realise*).

The body of the presentation will be organized in three main sections. The first is centred on explaining the theoretical indebtedness of our approach to lexical-constellation analysis, particularly to its revision of Mason's (2000) notion of "lexical gravity" (Cantos & Sánchez, 2001; Almela, 2011; Almela *et al.*, 2011, 2013). In this section we will also pinpoint the similarities and differences between the concept of "co-collocation" and the neighbouring notions of "second-order collocation" (Mollet *et al.*, 2011), "collocational network" (Williams, 1998, 2012), and "covarying collexeme analysis" (Stefanowitsch & Gries, 2005). The second section will provide a detailed description of the methodology. The third section will present the results obtained from applying this methodology to the analysis of nouns from the conceptual domain of CAUSALITY in the *enTenTen[2012]* corpus (accessed at SketchEngine).

References

Almela, M. 2011. Improving corpus-driven methods of semantic analysis: a case study of the collocational profile of 'incidence.' *English Studies,* 92(1): 84-99.

Almela, M., P. Cantos, & A. Sánchez. 2011. From collocation to meaning: revising corpus-based techniques of lexical semantic analysis. In I. Balteiro ed. *New Approaches to Specialized English Lexicology and Lexicography*. Newcastle upon Tyne: Cambridge Scholars Press, 47-62.

Almela, M., P. Cantos, & A. Sánchez. 2013. Collocation, co-collocation, constellation… Any advances in distributional semantics? *Procedia – Social and Behavioral Sciences,* 95: 231-40.

Cantos, P. & A Sánchez. 2001. Lexical constellations: what collocates fail to tell. *International Journal of Corpus Linguistics,* 6(2), 199-228.

Mason, O. 2000. Parameters of collocation: the word in the centre of gravity. In J. M. Kirk ed. *Corpora Galore. Analyses and Techniques in Describing English*. Amsterdam: Rodopi, 267-80.

Mollet, E., A. Wray & T. Fitzpatrick. 2011. Accessing second-order collocation through lexical co-occurrence networks. In T. Herbst, S. Faulhaber & P. Uhrig eds. *The Phraseological View of Language: A Tribute to John Sinclair*. Berlin: Walter de Gruyter, 87-121.

Stefanowitsch, A. & S. Th. Gries, 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory,* 1(1): 1-43.

Williams, G. 1998. Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics,* 3(1): 151-71.

Williams, G. 2012. Bringing the data and dictionary together: real science in real dictionaries. In A. Boulton, S. Carter-Thomas & E. Rowley-Jolivet eds. *Corpus-Informed Research and Learning in ESP. Issues and Applications* Amsterdam: Benjamins, 219-40.

Lexical *have*: When do you have the demeanour of a full verb, and when have you the behaviour of an auxiliary?

Laura Altenkirch & Florian Dolberg (Johannes Gutenberg-Universität Mainz)

The verb *have* can either function as an auxiliary verb marking perfective aspect (cf. 1) or as a lexical verb with a noun phrase complement indicating possession (cf. 2). This latter, lexical use of *have* also includes more abstract, metaphorical meanings such as relations (cf. 3) or health (cf. 4) (cf. e.g. Quirk et al. 1985: 131).

(1) […] many researchers **have** tended to underplay the problems that arose in the process of research. (BNC)

(2) Buyers of cheaper widgets now **have** money for other purchases. (BNC)

(3) He **has** a daughter of 23 by his first marriage. (BNC)

(4) I **have** a headache. (BNC)

Lexical *have* (cf. 2 through 4) exhibits a syntactic two-face nature: it mostly behaves like any other lexical verb but occasionally also like an auxiliary, even though sentence meaning and context clearly establish lexical meaning (i.e. possession) rather than auxiliary function (i.e. marking perfective aspect). Examples (5) and (6) illustrate this variation.

(5) Do you **have** any idea what the average annual rate is for this? (BNC)

(6) **Have** you any idea what that refers to? (BNC)

This variation not only pertains to questions, but also to negations as well as tag questions. While this phenomenon is frequently mentioned in reference grammars (cf. e.g. Biber et al. 1999: 160-163, 216; Huddleston & Pullum 2002: 113; Quirk et al. 1985: 131-132), the authors are in discord: Huddleston & Pullum assert that lexical *have* always behaves like a lexical verb in American English, whereas "in BrE the lexical use has become common too, and the auxiliary use is tending to sound relatively formal or old fashioned" (2002: 113). Biber et al. (1999: 161-163, 216) analysed the *Longman Spoken and Written English Corpus*, and found lexical *have* with auxiliary behaviour more frequently in British English than in American English. Counter to Huddleston & Pullum's (2002: 113) assertion, Biber et al. (1999: 161-163, 216) observed that American English does feature this variant occasionally. Outer-circle varieties, e.g. Indian, Hong Kong, and Singapore English, appear to differ from both the British and American template as regards lexical *have* (cf. Nelson 2004). While many studies concerned with the variable usage of lexical *have* contrast it with *have got* (cf. e.g. Tagliamonte et al. 2010), the intra-varietal distribution of auxiliary vs. full verb behaviour of lexical *have* appears largely unascertained, and the factors which facilitate or hinder lexical *have* to syntactically behave like an auxiliary are yet to be explored.

The present paper contributes to bridging this research gap by addressing the following questions:

- Is lexical *have* more frequently behaving like an auxiliary in colloquial, informal English, and, conversely, is lexical *have* chiefly behaving like a full verb in more formal registers?
- Are these stylistic differences stable across varieties?
- Are the distributional patterns the same in questions and negations, or does one context (dis)favour one of the variants more than the other?
- Is the choice between lexical *have* as a full verb or as an auxiliary furthermore influenced by pragmatic and/or cognitive factors, such as definiteness, animacy, and abstractness of possessor and possessum?

To approach these issues, the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) serve as databases for ascertaining the behaviour of lexical *have* in less formal English, represented by the sections Spoken and Fiction, in comparison to the behaviour of lexical *have* in more formal English, represented by the newspaper and academic texts featured in said corpora.

Preliminary results suggest that lexical *have* displays auxiliary behaviour more frequently in informal settings than in formal ones, counter to Huddleston & Pullum's (2002: 113) assertion. Moreover, we can tentatively infer at this point that American English in general appears less tolerant (but by no means intolerant) of lexical *have* behaving like an auxiliary, in keeping with Biber et al.'s (1999: 162) findings. The present paper empirically investigates whether lexical *have* behaving like an auxiliary is best characterised as a largely non-American, colloquial phenomenon and whether the syntactic variation of lexical *have* is further influenced by general pragmatic and basic cognitive predictors, such as definiteness, animacy, and abstractness of possessor and possessum.

References

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

Davies, M. 2004-. BYU-BNC. (Based on the British National Corpus, Oxford University Press. http://corpus.byu.edu/bnc/.

Davies, M. 2008-. The Corpus of Contemporary American English: 450 million words, 1990-present. http://corpus.byu.edu/coca/.

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

Nelson, G. 2004. Negation of lexical *have* in conversational English. *World Englishes*. (23.2): 299-304.

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

Tagliamonte, S. A., A. D'Arcy & B. Jankowski. 2010. Social work and linguistic systems: marking possession in Canadian English. *Language Variation and Change* (22): 149-73

GET, GET-constructions and the GET-passive in nineteenth-century English: Corpus analysis and prescriptive comments

Lieselotte Anderwald (University of Kiel)

Prescriptivism as a factor influencing (at least written) language is regularly invoked when corpus results are unexpected, but is rarely substantiated. In this talk, I will present one case study that raises interesting methodological and linguistic questions. The GET-passive, together with HAVE GOT, has seen considerable criticism by language mavens in the twentieth century (e.g. reported in Ballard 1939: 23-26; Mittins et al. 1970: 33-35). Since the rise of the GET-passive is a Late Modern English phenomenon (Denison 1993, 1998; Hundt 2001; cf. also Gries and Hilpert 2012 for the twentieth century), it makes intuitive sense to link its stigmatization with prescriptive grammar writing, which was demonstrably at its height in the nineteenth century. And indeed, prescriptive grammars have been blamed for the comparatively slow rise of the GET-passive over the course of the nineteenth century (Hundt 2001). However, this claim has not been supported by evidence from the grammars themselves yet, and it is not clear whether the evidence from the twentieth century can be projected back in time. For this reason, I will trace the rise of GET over the nineteenth (and into the twentieth) century corpus-linguistically in COHA (Davies 2010-) and differentiate the individual constructions that GET was (and is) used in. I will then correlate the shift in constructional use with prescriptive comments in nineteenth-century grammar books, based on my Collection of Nineteenth-Century Grammars, which contains over 250 grammar books from Britain and the US, and investigate temporal correlations (surely a prerequisite for causal connections). As my talk will show, criticism of GET and GET-constructions does occur in the nineteenth century; however, the GET-passive curiously is almost exempt from this criticism. Although it thus seems that criticism of the GET-passive is a truly twentieth-century phenomenon, the more general relegation of GET to an informal register can be linked to nineteenth-century prescriptive grammar writing, which may have contributed to the register-specificity of (all uses of) GET today.

References

Ballard, P. B. 1939. *Teaching and Testing English*. London: University of London Press.

Davies, M. 2010-. *The Corpus of Historical American English: 400 million words, 1810-2009*. http://corpus.byu.edu/coha/.

Denison, D. 1993. *English Historical Syntax: Verbal Constructions*. London & New York: Longman.

Denison, D. 1998. Syntax. In S. Romaine, ed. *The Cambridge History of the English Language*. Vol. 4. Cambridge: Cambridge University Press, 92-329.

Gries, S. Th., & M. Hilpert. 2012. Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics. In T. Nevalainen & E. Closs Traugott, eds. *The Oxford Handbook of the History of English*: *10*. Oxford: Oxford University Press, 134-44.

Hundt, M. 2001. What corpora can tell us about the grammaticalisation of voice in *get*-constructions. *Studies in Language*, 25: 49-88.

Mittins, W. H., M. Salu, M. Edminson & S. Coyne. 1970. *Attitudes to English Usage*. London: Oxford University Press.

To boldly split where almost everyone has split before! A corpus study of VP adverbial positions in American English

Sabine Arndt-Lappe (Trier University)

Despite its omnipresence in discussions of prescriptivist approaches to grammar, the so-called split infinitive construction has not received much attention in the linguistic literature on synchronic English syntax. Discussions in the descriptive literature are largely concerned with the grammatical wellformedness of the construction (e.g. Huddleston & Pullum eds. 2002: 575ff., 1018ff.), and the few empirical studies that are available provide indications that register and lexical biases are factors that influence usage of the construction (cf. esp. Fischer 2007, Calle Martín & Miranda-García 2009, Perales Escudero 2011). Quantitative studies, however, that test the potential of those factors to actually predict the occurrence of the split infinitive construction, are largely lacking. This is surprising especially in view of the fact that recent times have seen much scholarly interest in phenomena of grammatical variation in general, and in those that concern constituent ordering in particular (cf. e.g. Hawkins 1999, Wasow & Arnold 2003 on the relative ordering of adverbials; Bresnan et al. 2007, Bresnan & Ford 2010 on the dative alternation; Rosenbach 2005, Jäger & Rosenbach 2006, Szmrecsanyi 2010 on the genitive alternation). Pertinent work has not only provided key insights into the nature of the factors that determine speakers' choices between alternative constructions, but has also made available methodologies that capture the interaction of such factors in a multifactorial analysis of quantitative data.

The present paper addresses the issue of the split infinitive construction as a phenomenon of grammatical variation. A case study is presented that systematically investigates the placement of VP adverbials in *to* clauses sampled from the *Corpus of Contemporary American English* (COCA, Davies 2008-; N = 604). The findings do not only show that the split infinitive construction is frequent – they are also fully in line with much recent corpus-based and experimental empirical work on grammatical variation phenomena (cf. esp. Bresnan et al. 2007, Bresnan & Ford 2010, Gries & Stefanowitsch 2004, Wolk et al. 2013). I will argue that, like much of that work, the empirical findings of the present corpus study support a usage-based and probabilistic approach to grammar.

Thus, in a classification tree analysis, syntactic weight of the AdvP, but also of the object of the verb emerges as the strongest predictors of AdvP position. Interestingly, the effect is only partially compatible with traditional processing-based conceptions of weight effects as efficiency principles (cf. e.g. Hawkins 1994), lending support to the idea that processing effects may be constrained by (language-dependent) threshold levels (Wolk et al. 2013). Furthermore, the analysis identifies both semantic and lexical effects as significant correlates of AdvP placement in the corpus data. In particular, degree adverbs are particularly prone to appearing in the 'split infinitive position' whereas temporal adverbs are not. Also, there is a small but noticeable verb bias. Crucially, all effects found in the analysis are probabilistic. Unexpectedly, the factor 'register', which has been emphasised in much of the literature on the split infinitive construction, does not surface as a significant predictor in the present analysis.

References

Bresnan, J., A. Cueni, T. Nikitina & H. R. Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Kramer & J. Zwarts eds. *Cognitive Foundations of Interpretation.* Amsterdam: Royal Netherlands Academy of Science, 69-94.

Bresnan, J. & M. Ford. 2010. Predicting syntax: processing dative constructions in American and Australian varieties of English. *Language,* 86(1): 186-213.

Calle Martín, J. & A. Miranda-García. 2009. On the use of split infinitives in English. In A. Renouf & A. Kehoe eds. *Corpus Linguistics: Refinements and Reassessments*. Amsterdam: Rodopi, 347-64.

Davies, M. 2008-. *The Corpus of Contemporary American English: 450 million words, 1990-present*. http://corpus.byu.edu/coca/.

Fischer, R. 2007. To boldly split the infinitive - or not? In S. Elspaß, N. Langer, J. Scharloth & W. Vandenbussche eds. *Germanic Language Histories "from Below" (1700 - 2000).* Berlin: Walter de Gruyter, 259-74.

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

Hawkins, J. 1994. *A Performance Theory of Order and Constituency*. Cambridge: CUP.

Hawkins, J. 1999. The relative order of prepositional phrases in English: going beyond manner–place–time. *Language Variation and Change,* 11: 231-66.

Huddleston, R. & G. K. Pullum eds. 2002. *The Cambridge Grammar of the English Language*. Cambridge: CUP.

Perales-Escudero, M. 2011. To split or to not split: the split infinitive past and present. *Journal of English Linguistics,* 39(4): 313-34.

Rosenbach, A. 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81(3). 613–644.

Rosenbach, A. 2007. Emerging variation: determiner genitives and noun modifiers in English. *English Language and Linguistics,* 11(01): 143.

Szmrecsanyi, B. 2010. The English genitive alternation in a cognitive sociolinguistics perspective. In D. Geeraerts et al eds. *Advances in Cognitive Sociolinguistics.* Berlin: Mouton de Gruyter, 141-66.

Wasow, T. & J. Arnold. 2003. Post-verbal constituent ordering in English. In G. Rohdenburg ed. *Determinants of Grammatical Variation in English.* Berlin: Mouton de Gruyter, 119-54.

The roots of multi-word discourse markers: A corpus-based study of the *speaking of X* construction

Yinchun Bai (University of Freiburg / University of Antwerp)

Discourse markers have been investigated by different approaches, quite intensively in synchronic studies of their functions and morpho-syntactic features by way of discourse analysis (e.g. Ferrara 1997; Gohl and Günthner 1999; Fraser 2009), and to a lesser extent in studies of their developmental paths using diachronic data (e.g. Brinton 2008; Lewis 2011; Prevost 2011). This paper uses the Corpus of Historical American English (the COHA corpus) as data source and explores the possible input domains for word strings developing into multi-word discourse markers. It presents the case study of the *speaking of X* construction (further referred to as the Spox construction), as in *Speaking of L.A.'s freeways, how are you dealing with road rage?* (2009, MAG, COCA) and in *I'm perfectly serious. They wander down from Canada. Speaking of which, where did you wander from?* (2008, FIC, SourCherrySurprise). First, this paper offers a discussion of the formal and functional

properties of the Spox construction, supporting its recognition as a discourse marker. In a next step, it proposes three possible roots of motivation for change since the 19th century: (1) the co-occurring "*remind*" context – e.g. *Speaking of catching flies reminds me of political economy.* (1831, MAG, NewEngMag), (2) the sentence-initial adverbial – e.g. *Speaking of the general question of doing government work by contract, I expressed the view*[…] (1910, MAG, Scribners), and (3) the "*now that*" clause – e.g. *And now that we're speaking of profits, Mr. Crashly, I have this thought to put before you.* (1950, FIC, SomethingValue). It will be argued that each root contributes primarily, but not exclusively, to some aspect/s of the establishment and consolidation of the discourse marker status of the Spox construction. In conclusion, this paper shows that the development of the Spox construction as a multi-word discourse marker is rooted in the formal and conceptual blending of all three input domains.

References

Brinton, L. J. 2008. The Comment Clause in English: Syntactic Origins and Pragmatic Development. Cambridge: Cambridge University Press.

Fraser, B. 2009. Topic orientation markers. *Journal of Pragmatics*, 415: 892-98.

Ferrara, K. W. 1997. Form and function of the discourse marker anyway: implications for discourse analysis. *Linguistics*, 352. 343-78.

Gohl, C., & Günthner, S. 1999. Grammatikalisierung von weil als Diskursmarker in der gesprochenen Sprache. *Zeitschrift für Sprachwissenschaft*, 18(1): 39-75.

Lewis, D. M. 2011. A discourse-constructional approach to the emergence of discourse markers in English. *Linguistics*, 492: 415-43.

Prevost, S. 2011. A propos from verbal complement to discourse marker: a case of grammaticalization? *Linguistics*, 492, 391-413.

(Association) measure for measure: Comparing collocation dictionaries with co-occurrence data for a better understanding of the notion of collocation

Sabine Bartsch (TU Darmstadt)
Stefan Evert (FAU Erlangen-Nürnberg)
Thomas Proisl (FAU Erlangen-Nürnberg)
Peter Uhrig (FAU Erlangen-Nürnberg)

Lexical collocations are a complex phenomenon for which neither traditional nor cognitive linguistic theories have yet found satisfactory definitions that would allow for a lexicographically convincing operationalisation. Despite the pervasiveness of collocations in language and their importance for our understanding of the structure of human language as well as for many applications, their definition and characterisation leave many questions unanswered.

Corpus-based studies of collocation and the development of collocation extraction tools have been influenced by two principal views: (a) an empirical notion of collocation (Firth 1957), which builds upon the recurrent co-occurrence of lexical items in more or less clearly defined

contexts; (b) phraseological notions of collocation which are prevalent in lexicography and characterise collocations on the basis of their semantic, syntactic and distributional irregularity (cf. Hausmann 1979; 1984; 1985; Manning & Schütze 1999: 184). Other, related definitions equate collocations with lexicalised multiword expressions (as is often done in computational linguistics, e.g. Choueka 1988) or focus on their cognitive reality, using evidence from priming studies (Durant & Doherty 2010) or plausibility judgements (Lapata *et al.* 1999).

The operationalisation of such collocation definitions, which is necessary to allow for their reliable identification in corpora, remains a notoriously difficult issue. The situation is similar for related questions such as the choice of an appropriate quantitative measure of the association between co-occurring words, the influence of the quality and size of the corpus, and the qualitative evaluation of automatically extracted collocation candidates.

The aim of this paper is to report on work towards a better understanding of different notions of collocation and their operationalization and towards gauging the reliability of automatic collocation identification in large corpora. To this end, the research reported in this paper compares a sample inventory of collocations listed in two specialized collocation dictionaries (the pre-corpus era BBI and the corpus-based OCD2; see also Lea 2007) with measures of statistical association in linguistic corpora of different sizes and with different levels of linguistic pre-processing. The resulting collocation candidates are manually evaluated against a well-defined subset of data from the two dictionaries.

It will be shown that at least for common general language collocations such as those listed in dictionaries, smaller and cleaner corpora such as the BNC deliver better lists of collocation candidates than larger, but noisy web corpora. Furthermore it will be shown that syntactically annotated data are not only superior for collocation extraction from BNC-like corpora (as shown in previous work), but also for web-based corpora despite the relatively low accuracy of automatic syntactic annotation tools on such data.

Comparing different statistical association measures – such as log-likelihood ratio, t-score, chi-squared, several variants of Mutual Information and directional measures such as ΔP – with the dictionary data, we discover some surprising facts: MI^2 (Daille 1994: 193) and t-score correspond better to lexicographers' intuitions than the classical MI measure that has long been popular in computational lexicography; the widely-held belief that the chi-squared test is unsuitable for collocation identification (Dunning 1993) is not always true; finally, the relative usefulness of an association measure depends much less on the quality, amount and annotation of the corpus data than on the particular notion of collocation to be identified (i.e. on which dictionary is used as a "gold standard").

Thus – returning to the original question as to the concept of collocation – we can show that the collocation dictionaries used in the present study differ substantially with respect to their view of what should be listed as a collocation, which may (at least in part) be due to the fact that one of the two was created in the pre-corpus era. The automatic evaluation allows us to compare both dictionaries against the corpus findings but also to compare to what extent the

explicit definition of collocation (as stated by the editors) and the implicit definition (i.e. the selection of collocations) correspond.

References

BBI: *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*. Amsterdam: Benjamins, 1986. [3rd ed 2010.]

Choueka, Y. 1988. Looking for needles in a haystack. In *Proceedings of Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications): RIAO "88*, Cambridge, MA, 609-23.

Daille, B. 1994. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7. http://www.bdaille.com/index.php?option=com_docman&task=doc_download&gid=8. Accessed 27.03.2015.

Dunning, T. E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.

Durrant, P. & A. Doherty. 2010. Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6(2), 125-55.

Firth, J. R. 1957. Papers in Linguistics 1934–1951. Oxford: OUP.

Hausmann, F. J. 1979. Un dictionnaire de collocations est-il possible? *Travaux de linguistique et de littérature*, 17. 187-95.

Hausmann, F. J. 1984. Wortschatzlernen ist Kollokationslernen. *Praxis des neusprachlichen Unterrichts*, 31. 395-406.

Hausmann, F. J. 1985. Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholtz & J. Mugdan eds. *Lexikographie und Grammatik*. Tübingen: Niemeyer, 118-29.

Lapata, M., S. McDonald & F. Keller. 1999. Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999)*, Bergen, Norway, 30-36.

Lea, D. 2007. Making a collocations dictionary. *Zeitschrift für Anglistik und Amerikanistik*, 55(3), 261-72.

Manning, C. D. & H. Schütze 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA/London: MIT Press.

OCD2: *Oxford Collocations Dictionary for Students of English*. 2nd ed. Oxford, 2009.

From light verb constructions to negative polarity items: The cases of *take notice of* and *make mention of*

Eva Berlage (University of Hamburg)

It is well known that English has a series of so-called negative polarity items (NPIs) whose occurrence is restricted to or strongly preferred in non-assertive contexts. Typical examples include the *any* class of items (e.g. *any*, *anybody*, *any longer*, *any more*, *anything*), various grammatical items (e.g. *much*, *either*, *ever*, *yet*), the modal auxiliaries *dare* and *need*, a few lexical verbs (e.g. *bother* + infinitival, *budge*, *faze*) and a vast range of idioms such as *can be bothered*, *give a damn*, *see a (living) soul* etc. (for a more comprehensive list, see e.g. Huddleston/Pullum et al. 2002: 823; von Bergen/von Bergen 1993). Far less well known is the evolution of these polarity sensitive items and the fact that the development of some of them can be linked to such historical processes as lexicalisation or its counter-image (for the

process of lexicalisation and its counter-image, see e.g. Brinton/Traugott 2005; Berlage 2010, 2014; Traugott/Trousdale 2013).

In this paper, I will focus on the evolution of the constructions *take notice of* and *make mention of*, which started off as light verb constructions and, in the course of time, have developed a second function as negative polarity items. Using million-word corpora of fictional British texts from 1600 to the present-day (based on the Chadwyck-Healey collection of prose texts spanning the 16th to 19th centuries and the fictional components of the BNC), I will attest to the increasing co-occurrence of the nominal in these constructions with the determiners *any* and *no* as in (1), (2) and (5).

(1) […] the Abbess wondered yet more, but did **not** seem to take any notice of the Horns, (Early English Prose Fiction Corpus, 1682)
(2) The writer made **no** mention of returning to town; (Nineteenth Century Fiction Corpus, 1811)
(3) **Do you mind** about my taking notice of it? (Nineteenth Century Fiction Corpus, 1866)
(4) […] who dragged books from shelves, and **if** he took notice of a lap-dog only forced open its mouth to look at his teeth (Nineteenth Century Fiction Corpus, 1862)
(5) It is true, that on seeing him there she might have **forbore** making any mention of Belpine, (Eighteenth Century Fiction Corpus, 1753)

My paper provides a qualitative and quantitative discussion of how strong a non-assertive context needs to be for *take notice of* and *make mention of* to occur and asks whether the preference of these NPIs either for strictly negative contexts as in (1) and (2) or for other non-assertive contexts as in (3)-(5) has changed in the course of time. My research thereby adds to the rich theoretical and empirical literature on the question of what licenses the occurrence of NPIs (cf. e.g. Ladusaw 1980; Fauconnier 1975; 1979; Kadmon/Landman 1993; Chemla et al. 2011; Israel 1996; 2011). On the basis of my empirical findings, I argue that we have to take into account the contextual information that is available from the sentence or the wider utterance that an NPI occurs in if we want to predict its distribution.

The other major theoretical implication of my paper is for discussions of lexicalisation: how does it come about that items which tend to be relatively fixed at one point (rarely allowing for material to occur between verb (e.g. *take*) and complement (e.g. *notice*)) become syntactically more flexible and continuously allow for the insertion of grammatical material (e.g. the determiners *no* and *any*)? Given the increasing co-occurrence of the nominal with such determiners, would these constructions still qualify as instances of lexicalisation? On the basis of the empirical data presented my paper aims to enrich the discussion of lexicalisation and what we may refer to as ‘delexicalisation’.

References

Berlage, E. 2010. The lexicalisation of predicative complements in English. *Transactions of the Philological Society* 108 (1): 53-67.

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: Cambridge University Press, 207-23.
Brinton, L. & E. Closs Traugott eds. 2005. *Lexicalization and Language Change*. Cambridge: Cambridge University Press.
Chemla, E., V. Homer & D. Rothschild. 2011. Modularity and intuitions in formal semantics: the case of polarity items. *Linguistics and Philosophy* 34: 537-70.
Fauconnier, G. 1975. Polarity and the scale principle. In R. E. Grossman, L. J. San & T. J. Vance eds. *Proceedings of the 11th meeting of the Chicago Linguistics Society*. Chicago: CLS. 188-99.
Fauconnier, G. 1979. Implication reversal in natural language. In F. Guenther & S. J. Schmidt eds. *Formal Semantics and Pragmatics for Natural Languages*. Dordrecht: Reidel, 289-301.
Huddleston, R., G. K. Pullum et al. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Israel, M. 1996. Polarity sensitivity as lexical semantics. *Linguistics and Philosophy* 19: 619-66.
Israel, M. 2011. *The Grammar of Polarity: Pragmatics, Sensitivity, and the Logic of Scales*. Cambridge: Cambridge University Press.
Kadmon, N. & F. Landman. 1993. Any. *Linguistics and Philosophy* 16: 353-422.
Ladusaw, W. 1980. *Polarity Sensitivity as Inherent Scope Relations*. New York/London: Garland Publishing.
Traugott, E. Closs & G. Trousdale. 2013. *Constructionalization and Constructional Changes.* Oxford: Oxford University Press.
von Bergen, A. & K. von Bergen. 1993. *Negative Polarität im Englischen*. Tübingen: Narr.

Identifying linguistic epicentres empirically: The case of South Asian Englishes

Tobias Bernaisch (Justus Liebig University Giessen)
Stefan Th. Gries (University of California, Santa Barbara / Justus Liebig University Giessen)

A linguistic epicentre can generally be identified on the basis of two criteria, namely "if it shows endonormative stabilization (i.e. widespread use, general acceptance and codification of the local norms of English)[...] on the one hand, and the potential to serve as a model of English for (neighbouring?) countries on the other hand" (Hundt 2013: 185). Along these lines, "(pre-) epicentric influence" (Peters 2009: 122) has been traced from Australian on New Zealand English and Leitner (cf. 1992: 225) posits that India is a linguistic epicentre for South Asia. Given that studies of epicentral variety constellations, however, "still lack the empirical evidence that would allow us to make more than educated guesses" (Hundt 2013: 186), this study suggests an empirical, corpus-based method of epicentre identification and applies it to South Asian Englishes.

With a focus on the dative alternation, i.e. the alternation between the double-object construction (e.g. *John gave Mary a book.*) and the prepositional dative (e.g. *John gave a book to Mary.*), the present paper studies the norms underlying this constructional choice in six South Asian Englishes and British English. The corpus data for South Asia stem from the 18m-word-large *South Asian Varieties of English* (SAVE) *Corpus* (cf. Bernaisch 2011) sampling variety-specific acrolectal newspapers texts of Bangladeshi, Indian, Maldivian, Nepali, Pakistani and Sri Lankan English and the newspaper texts in the *British National*

Corpus are used as British English equivalents. Via *Multifactorial Prediction and Deviation Analysis with Regression* (MuPDAR; cf. e.g. Gries & Adelman (2014)) under consideration of nested random effects, we a) identify the factors that cause South Asian speakers of English to make constructional choices different from British English speakers when other influences on the constructional choice are controlled for and b) the South Asian linguistic epicentre in a completely bottom-up fashion by empirically validating the linguistic model which best represents the norms underlying the dative alternation in South Asian Englishes.

1381 examples were – under consideration of earlier findings on the dative alternation (cf. e.g. Gries 2003, Bresnan & Hay 2008, Schilk et al. 2013, Bernaisch et al. 2014) – annotated for eight syntactic (e.g. length of patient and recipient), semantic (e.g. the semantic class of the ditransitive verb), pragmatic (e.g. the discourse accessibility of patient and recipient) and data-structure-related variables (e.g. the newspaper from which a given example was taken). In terms of differences between British English and South Asian speakers of English, the results *inter alia* show that speakers of South Asian Englishes choose more prepositional datives than British English speakers when the patient or the recipient are not introduced in the preceding discourse and when there is no marked difference in the lengths of patient and recipient. Consequently, discourse accessibility of patient and recipient seems to be an actuator of structural nativisation (cf. Schneider 2003, 2007) in South Asian Englishes. Based on how well a variety-specific model derived via MuPDAR could predict constructional choices in the remaining varieties, we are able to show that it is valid to assume that Indian English functions as a linguistic epicentre for South Asia – at least in relation to the dative alternation. Given that Indian English can be regarded as an endonormatively stabilised variety (cf. Mukherjee 2007: 163), this finding is certainly in accordance with the advanced evolutionary status linguistic epicentres should theoretically display (cf. Hundt 2013: 185) and provides strictly empirical evidence for earlier, partly introspective perspectives on epicentral configurations in South Asia (cf. e.g. Leitner 1992).

References

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

Bernaisch, T., S. Th. Gries & J. Mukherjee. 2014. The dative alternation in South Asian Englishes. modelling predictors and predicting prototypes. *English World-Wide*, 351: 7–31.

Bresnan, J. & J. Hay. 2008. Gradient grammar: an effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua*, 118: 245–59.

Gries, S. Th. 2003. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, 1: 1–27.

Gries, S. Th. & A. S. Adelman. 2014. Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research. In Jesús Romero-Trillo, ed. *Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*. Cham: Springer, 35–54.

Hundt, M. 2013. The diversification of English: old, new and emerging epicentres. In D. Schreier & M. Hundt, eds. *English as a Contact Language*. Cambridge: Cambridge University Press, 182–203.

Leitner, G. 1992. English as a pluricentric language. In M. Clyne, ed. *Pluricentric Languages: Differing Norms in Different Nations*. Berlin: Mouton de Gruyter, 179–237.

Mukherjee, J. 2007. Steady states in the evolution of New Englishes: present-day Indian English as an equilibrium. *Journal of English Linguistics*, 352. 157–187.
Peters, P. 2009. Australian English as a regional epicentre. In T. Hoffmann & L. Siebers, eds. *World Englishes – Problems, Properties and Prospects*. Amsterdam/Philadelphia: Benjamins, 107–24.
Schilk, M., J. Mukherjee, C. F.H. Nam & S. Mukherjee. 2013. Complementation of ditransitive verbs in South Asian Englishes: a multifactorial analysis. *Corpus Linguistics and Linguistic Theory*, 92: 187–225.
Schneider, E. W. 2003. The dynamics of New Englishes: from identity construction to dialect birth. *Language*, 792: 233–281.
Schneider, E. W. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press.

Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus

Yves Bestgen (Centre for English Corpus Linguistics)
Sylviane Granger (University of Louvain)

In the last decade learner corpus research has been characterized by a wide range of studies focused on phraseological aspects of learner language, in particular collocations (Nesselhauf 2005) and lexical bundles (Chen & Baker 2010). Most of these studies are cross-sectional, i.e. they compare learner populations at a particular point in time. Due to the lack of longitudinal corpora, very few studies have tracked the phraseological development of the same learners over an extended period of time. Fortunately, this is now beginning to change thanks to recent corpus collection initiatives launched with a view to filling that gap. In our presentation we will discuss the results of an investigation of the learner phrasicon based on the *Longitudinal Database of Learner English*[1] *(LONGDALE)*, a collection of spoken and written data produced by EFL learners over a period of three years. For our study we made use of a subcorpus consisting of 178 texts written by 89 French-speaking learners and containing approximately 100,000 words in total. Each learner contributed an argumentative essay in their first year at university (Y1) and another in their third year (Y3); the topics of the essays were the same in Y1 and Y3 to ensure maximum comparability.

To assess learners' phraseological development, we made use of a methodology relying on *collgrams*, i.e. word bigrams that have been assigned an association score on the basis of a large reference corpus. Although there is already an abundance of terms to refer to phraseological units, we introduce this new term to avoid all confusion with *collocations*, which refer to combinations of significantly associated words which may but need not be contiguous, and *n-grams*, which refer to contiguous word sequences whose strength of association is unspecified. The method draws its inspiration from a study by Durrant & Schmitt (2009), who found that learners tend to overuse high-frequency collocations (identified on the basis of t-score) and underuse lower-frequency, but strongly associated, collocations (identified by their mutual information (MI)). The objective of our study is to find out whether a similar trend can be established on the basis of longitudinal learner data.

[1] http://www.uclouvain.be/en-cecl-longdale.html

The collgram method involves four successive steps: (1) extraction of four categories of bigrams from each text of the learner corpus: all the bigrams whatever their part-of-speech and three syntactically-defined bigram categories: noun-noun, adjective-noun and adverb-adjective; (2) assignment of two association scores (MI and t-score) to each bigram on the basis of the *British National Corpus*; (3) classification of the bigrams as either absent from the reference corpus or displaying one of four degrees of collocational strength; (4) computation of the means of three measures (MI, t-score and proportion of bigrams absent from the reference corpus) for both types and tokens.

Using randomization tests for matched samples (Howell 2013), we found a large number of significant differences between Y1 and Y3 essays. Although the results differ somewhat according to the category of bigram, there is a general tendency for Y3 texts to contain fewer non-collocational bigrams (*past people, relevant in, more little*) and fewer high-scoring t-score collgrams (*in particular, close to, was going*), but more high-scoring MI bigrams (*paramount importance, funnily enough, life expectancy*). The results are very similar to those obtained when the same method is applied to cross-sectional data (Granger & Bestgen 2014). Overall, the results therefore suggest that the collgram method is capable of discriminating learners at different proficiency levels as well as tracking their development over the course of the learning process.

In our conclusion we will point to some applications of the method for teaching, testing and automated scoring.

References

Chen, Y.-H. & Baker, P. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14, 2: 30-49.

Durrant, P. & Schmitt, N. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching IRAL* 47: 157-77.

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 IRAL* 523: 229-52.

Howell, D. 2013. *Statistical Methods for Psychology*. 8th ed. Wadsworth, Cengage Learning.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: Benjamins.

GET and GET-PVs in World Englishes

Elisabeth Bruckmaier (LMU Munich)

GET is a highly frequent and multifunctional English verb. This analysis sets out to shed light on its use in the language systems of British English, Jamaican English, and Singaporean English. The data used come from the Lancaster-Oslo/Bergen Corpus (LOB), the Freiburg-LOB Corpus (FLOB), and three corpora of the International Corpus of English (ICE) project: ICE-Great Britain, ICE-Jamaica, and ICE-Singapore. While LOB and FLOB are used for making statements about the diachronic development of GET in British English, the ICE

corpora allow a broadening of the picture to World Englishes: ICE-Great Britain has been chosen as a background of comparison, with English used as a native language, and ICE-Jamaica and ICE-Singapore are representatives of English as a second language. While in Jamaica, the creole continuum constitutes a special sociolinguistic situation, the English used in Singapore is said to be relatively advanced in terms of structural nativisation and gradually becoming a first language.

Corpus-driven and corpus-based approaches complement each other. While theory-free induction is hardly possible, I try to let the data speak for themselves: categories have emerged in the course of the analysis of the data and are seen as dynamic rather than as fixed classes, often giving place to continua of description.

A description of the token frequencies of the individual word-forms of GET will introduce the analysis. Looking at simple token frequencies of GET is an interesting enterprise in itself, the reason being that few other verbs have attracted as much prescriptive resistance as GET. The question is whether, in the case of GET, particular patterns or uses can be linked to particular word-forms, which would support Sinclair's (1991) plea for paying heed to the word-form. Most researchers have so far glossed over differences between inflectional forms and have either lumped together different word-forms under the heading of lemma or have left the actual word-form that was the object of study unspecified or used the infinitive form (cf. Gries 2011). In fact, uneven distributions of word-forms of GET across ICE, particularly of *got* and *gotten*, indicate strong variability in World Englishes.

As a case study, results from an analysis of particle verbs in which GET occurs (GET-PVs) will be presented. Frequencies of GET-PVs are shown to be stable in diachrony in written British English. ICE-Great Britain displays the widest range of types as well as the largest pool of meanings in which GET-PVs are used. The Postcolonial Englishes analysed show much lower frequencies of GET-PVs but a higher percentage of simple types of PVs than British English, with complexity and cognitive processing time adduced as factors hampering the use of GET-PVs in the Postcolonial Englishes.

References

Gries, S. Th. 2011. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? in M. Brdar, S. Th. Gries & M. Žic Fuchs, eds. *Cognitive Linguistics: Convergence and Expansion.* Amsterdam/Philadelphia: Benjamins, 237-56.

Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation.* Oxford: OUP.

On reading late modern intentions: A corpus-based analysis of the late modern English subjectification of *be going to*

Sara Budts (KU Leuven)
Peter Petré (Université Lille 3, UMR 8163 / KU Leuven)

In this talk we provide a detailed account of the formal, functional and semantic changes that *be going to* underwent in Late Modern English. We know the history of *be going to*, a classic of grammaticalization studies (see Traugott 2012 for a recent overview), in quite some detail. Specifically for the late modern period, the period of *be going to*'s maturation, Disney (2009) hypothesises that *be going to* extended from encoding intention to encoding prediction through increased guessing of other people's intentions, resulting in its evidential semantics. While invaluable, Disney's study is based on a limited corpus, and mostly qualitative in nature. We test and refine Disney's hypothesis, and integrate it into the theory of subjectification (in the sense of Traugott 1989), on the basis of a large-scale corpus study. Despite *be going to*'s popularity, such studies remain rare. A recent large-scale study has focused on the initial stages of this process (Petré & Van de Velde 2014), but similar studies are still lacking for the late modern period. Hilpert (2008) provides some insight in the extension to new types of infinitive, but does not really go into details regarding the underlying mechanisms of change.

Data are based on an extensive sample from CLMETEV (De Smet, Diller & Tyrkkö 2011) complemented by exhaustive data from PPCMBE (Kroch, Santorini & Diertani 2010). In total, 1257 attestations were analysed on quantifiable, mostly formal features at various levels of the construction, such as contraction of *be* (substantive level), extension to new subject types (e.g. raised, empty subjects) and types of infinitives (e.g., states rather than actions) (schematic level), and the sentence type and presence of epistemic and evidential marking (level of the host clause).

The corpus data reveal that the shift from intention to prediction started to take place in the first half of the eighteenth century, and originated in contexts with third person subjects, often in past tense narratives. Naturally, reporting the intention of others generally involves a certain amount of guesswork. This amount increased by the middle of the eighteenth century, when *be going to* started to occur with non-imminent infinitival complements. This naturally resulted in an additional, epistemic layer of prediction that gradually gained strength during the second half of the eighteenth and the beginning of the nineteenth century, allowed the underlying meaning of intention to wither, and the construction to become more and more frequent in the present tense.

The shift from intention to prediction is arguably an increase in subjectivity, as the emphasis gradually moved away from the grammatical subject to the speaker: what mattered was no longer the intentions of the subject, but the knowledge of the speaker about them. Traugott (1989) shows that other English epistemic auxiliaries such as *will* and *shall* go through a similar stage of subjectification. Interestingly, while there are significant differences between *will*, *shall*, and *be going to*, it appears that each goes through an intermediary stage that involved past tense uses with reference to a future in the past, which was already known to

the speaker. From a theoretical point of view, our analysis provides further evidence for recent claims (e.g. De Smet 2012) that grammaticalization follows minimally disruptive pathways, taking the smallest steps possible in the development, and at the same time shows that recurring patterns may be found at this smallest level as well.

References

De Smet, H. 2012. The course of actualization. *Language*, 88 (3), 601-33.

De Smet, H., H.-J. Diller & J. Tyrkkö. 2011. *The Corpus of Late Modern English Texts*, version 3.0. https://perswww.kuleuven.be/~u0044428/clmet3_0.htm.

Disney, S. 2009b. The grammaticalisation of *be going to*. *Newcastle Working Papers in Linguistics*, 15: 63-82.

Hilpert, M. 2008. *Germanic Future Constructions: A Usage-Based Approach to Language Change*. Amsterdam: Benjamins.

Kroch, A., B. Santorini & A. Diertani. 2010. *Penn Parsed Corpus of Modern British English*. http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html.

Petré, P. & F. Van de Velde. 2014. Tracing real-life agents' individual progress in ongoing grammaticalization. *How grammaticalization processes create grammar: from historical corpus data to agent-based models*. EvoLang. Vienna, 14-17 July 2014, 48-51.

Traugott, E. Closs. 1989. On the rise of epistemic meanings in English: an example of subjectification in semantic change. *Language*, 65: 31-53.

Traugott, E. Closs. 2012. On the persistence of ambiguous linguistic context over time: implications for corpus research on micro-changes.In J. Mukherjee & M. Huber eds. *Corpus Linguistics and Variation in English*: Theory and Description. Amsterdam: Rodopi, 231-46.

Tracking a change in progress: *Any-* and *no-*negation in spoken corpora of British and Canadian English

Claire Childs (Newcastle University)
Christopher Harvey (University of Toronto)
Karen Corrigan (Newcastle University)
Sali Tagliamonte (University of Toronto)

The expression of negation with indefinites in English is highly variable (Tottie 1991). It can be expressed on the verb alongside an indefinite *any-* pronoun (1) or can combine with the indefinite to result in a *no-* form, as in (2). Another possibility is (3), negative concord. These variants persist across varieties of English as a result of longitudinal grammatical change and social influences.

(1) My parents had*n't any* money.

(2) a. He had *no* money at all.
b. 'Cause there was *nothing* like child benefit.

(3) We did*n't* know *nothing* about this.

In Old English, the particle *ne* was key to the expression of negation, but by the Middle English period, the system had developed two possibilities. One favoured a negative indefinite pronoun after the verb (2) and the other involved negative concord (3). The social changes of the Early Modern period led to the middle classes eschewing negative concord and appropriating type (1) (Nevalainen 1998: 275). However, there has been competition

between all three variants ever since, suggesting that the dialectic between social evaluation and linguistic change has not yet been resolved. This raises a number of questions: What is the current state of this variability in Britain where it evolved and in North America where it was transported to? What can socially stratified corpora and a cross-variety perspective tell us about this case of linguistic variation and change?

Our data comprise socially stratified, community corpora of vernacular speech from research projects spanning Britain and Canada (1997-2010). Despite the infrequent occurrence of negative sentences generally, these corpora are large enough to analyse nearly 3000 tokens (Britain N=1204; Canada N=1765) and code them for factors previously reported to influence variant choice, including verb type and traditional social factors like age, sex and education. A comparative sociolinguistic approach using distributional analyses as well as fixed and mixed effects statistical modelling (Team 2007) reveals that *no*-negation is stoutly retained in Britain (75%) but remains a minority form in Canada (46%). Linguistic constraints hold cross-dialectally: main verb *BE* and *HAVE* retain *no*-negation, while lexical verbs favour *any*. However, the social embedding is community-specific. In Britain, sex effects are discernible as is an apparent time decline in *no*-negation in the North East, but there is age grading in York. In Canada, education effects outside large urban centres are apparent. The unique perspective gained from analysing comparable, socially stratified spoken corpora enables us to disentangle the influence of linguistic versus social factors. While the British and Canadian communities share a common variable grammar, the social value in choosing one variant over the other is highly localised.

References

Nevalainen, T. 1998. Social mobility and the decline of multiple negation in Early Modern English. In J. Fisiak & M. Krygier eds. *Advances in English Historical Linguistics (1996).* Berlin: Mouton de Gruyter, 263-91.

Team, R Development Core. 2007. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Tottie, G. 1991. *Negation in English Speech and Writing: A Study in Variation.* London: Academic Press.

A *not uninteresting* construction: Distributions and functions of double-negated adjectives

Claudia Claridge (University of Duisburg-Essen)

Expressions like *not unusual*, *not unimportant*, i.e.double-negated adjectives, have been given some attention in semantics and pragmatics (e.g. Jespersen 1917, Bolinger 1972, Horn 1989, 1991, 2010), but have so far not been investigated in actual language use at all. The primary aim of this paper is to investigate frequencies, distributions and functions of such forms with the help of the BNC. The second aim is to explore the potential of construction grammar to shed further light on these forms.

The focus in data analysis lies on *not un*-A(djective), as opposed to forms with the prefixes *in-*/*im-*/*irr-*/*ill-* or *dis-*, as *un*-types provide 70% of all hits and may thus be seen as representative of the whole construction. The other prefixal forms will still be taken into account whenever generalisations across prefixes seem promising. Also, adverbial *uncommonly*, *unduly*, although parallel to the adjectival cases, will mostly be ignored for the time being (the majority, c. 90%, of *un*-hits are adjectival). This leaves 1,681 hits to investigate in detail.

Obviously, *not un*-A is not very frequent overall, which goes along with the marked nature of this double-negated form (cf. Levinson 2000, Fraenkel & Schul 2008). It is found to occur more often in written contexts generally, and there in factual and scientific writing especially. Within spoken contexts, it prefers monologue over dialogue and somewhat more formal (so-called context-governed) contexts over casual conversation. Possible reasons for these distributions will be discussed.

In the overwhelming majority (80-90%), *not un*-A is used in predicative constructions, which puts it potentially, though not necessarily, in positions of syntactic prominence and informational focus – it may thus be used often to convey 'new' information. A closer contextual analysis will need to clarify this. The construction co-occurs in the present data with only 199 different adjectives, many of which appear only once, whereas four adjectives (*uncommon, unusual, unknown, unreasonable*) account for more than half of all tokens. The raw figures also indicate that adjectives from the semantic field(s) of frequency and typicality are attracted to this construction (ca. 50% of occurrences).

With regard to the construction's function, Horn (1991, 2010) provided a list of seven motives for its use, which, however, is not based on a solid empirical foundation. A preliminary inspection of parts of the data found frequent evidence for three items of his list, namely semantic mitigation, pragmatic mitigation (politeness), and contradiction. The latter is here used in a somewhat extended meaning compared to Horn's, taking not only contradiction to immediate context into account but also to common/implicit assumptions (cf. Givón 1993, Israel 2004 regarding negation and the relevance maxim). Clearly litotic uses, although variously mentioned in the literature (e.g. Bolinger 1972, Levinson 2000), seem to be rare.

A collostructional analysis (e.g. Stefanowitsch & Gries 2003, Hampe 2011) will be carried out to shed light on the interaction of the perceived syntactic, lexical and semantic-pragmatic tendencies – and answer the question of whether this is a/one construction or several constructions.

References

Bolinger, D. 1972. *Degree Words.* The Hague: Mouton.

Fraenkel, T. & Y. Schul. 2008. The meaning of negated adjectives. *Intercultural Pragmatics,* 5-4: 517–40.

Givón, T. 1993. *English Grammar.* Amsterdam: Benjamins.

Hampe, B. 2011. Discovering constructions by means of collostruction analysis: the English denominative construction. *Cognitive Linguistics* 22 (2): 211–45.

Horn, L. 1989. *A Natural History of Negation*. Chicago/London: U of Chicago Press.
Horn, L. 1991. *Duplex negatio affirmat*…: The economy of double negation. In L. M. Dobrin, L. Nichols & R. M. Rodriguez eds.. *Papers from the 27th Regional Meeting of the Chicago Linguistic Society. Part 2: The Parasession on Negation*. 80-106.
Horn, L. 2010. Multiple negation in English and other languages. In L. R. Horn ed. *The Expression of Negation*. Berlin: Mouton de Gruyter, 111-48.
Israel, M. 2004. The pragmatics of polarity. In L. R. Horn & G. Ward eds. *The Handbook of Pragmatics.* Oxford: Blackwell, 701-23.
Jespersen, O. 1966[2] [1917]. *Negation in English and Other Languages*. Copenhagen: Høst.
Levinson, S. 2000. *Presumptive Meanings*. Cambridge, Mass.: MIT Press.
Stefanowitsch, A. & S. Th. Gries. 2003. Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8 (2): 209-43.

New Englishes in clusters: Profiling verb complementation constructions across post-colonial Englishes

Sandra C. Deshors (New Mexico State University)

The question of why speakers match a particular complement type to a given predicate is not trivial. Predicates and complements combine in systematic and meaningful ways (Noonan 1985) and complementation is never random (Bourke 2007). This study explores verb complementation patterns by contrasting gerundial and infinitival constructions (*Marcus started drawing a picture* vs. *Marcus started to draw a picture*) across native English and four post-colonial English (ESL) varieties. Following the usage-based tradition, the two constructions are approached as composites of semantic and morpho-syntactic features. Ultimately, this study

(i) identifies the semantic and morpho-syntactic features intrinsic to ing- and to-complement constructions, and

(ii) assesses to what extent the distribution of those co-occurring grammatical features varies within complement constructions across ESL varieties.

From a usage-based perspective, "[g]rammar resides in patterns of composition which take the form of constructional schemas". These patterns sanction "the progressive assembly of expressions of any size and degree of symbolic complexity" (Langacker 2000: 20). Using verb complementation constructions therefore requires ESL speakers to acquire and represent statistical information about the distribution of linguistic features and to recognize the systematic variation in those features' distribution/combination across *-ing* and *-to* complement constructions. Because within constructions linguistic features compete with one another and because cross-linguistically speakers with different native linguistic backgrounds assign those features varying degrees of strength, it is a challenge for ESL speakers to produce native-like complement patterns.

Recent corpus-based ESL studies show that variation in the use of syntactic constructions by ESL speakers is best captured using multifactorial methodological approaches that cut across the semantic and morpho-syntactic levels. For instance, with this approach Nam et al. (2013)

have unveiled the governing principles behind ESL speakers' choices of the ditransitive, prepositional dative and monotransitive constructions of GIVE. The present study extends existing multifactorial work by combining two sophisticated statistical methods, logistic regression (to identify the semantic and morpho-syntactic features intrinsic to *ing*- and *to*-constructions) and cluster analysis (to assess to what extent *ing/to* constructions across ESL varieties differ with regard to those intrinsic features). With this two-step methodological approach, I analyze over 7700 occurrences of *-ing* and *to*- verb complements across native English and Hong Kong, Indian, Singaporean and Jamaican Englishes as extracted from the *International Corpus of English* and annotated for thirty-nine grammatical features (e.g., form of the object, semantic of complement's/matrix's lexical verb, etc.).

The regression results show that five linguistic features are intrinsic to *ing*- and *to*-constructions: the type and the semantics of predicated verbs, the semantics of complement verbs, voice in the predicated clause and the form in which the object is expressed (likelihood ratio = 303.44; $p < {}^{-52;}$ R^2 = 0.30; 84.5% classification accuracy; C = 0.81). Across ESL varieties, gerundial and infinitival constructions reflect much plasticity. The cluster analysis reveals that different ESL populations assign different weights to individual lexico-grammatical elements within complement constructions, empirically supporting the usage-based notion that when a complex structure coalesces into a unit, its subparts do not thereby cease to exist (Langacker 1987). For instance, unlike Hong Kong and Jamaican English speakers, Singaporean English speakers are shown to associate the predictor Voice with gerundial complements in ways that are most nativelike. Overall, the cluster results uncover a wide range of partial sanctioning structures (i.e., structures demonstrating a relationship of extension and conflict in specification) characteristic of the ESL varieties. Ultimately, the present study provides empirical evidence showing how native and ESL speakers differ in their knowledge of constructional patterns. As the first corpus study to assess within-construction variation across English varieties, this study unveils new patterns of variation across post-colonial Englishes.

References

Bourke, J. M. 2007. Verbal complementation: a pedagogical challenge. *Reflections on English Language Teaching*, 6(1): 35-50.

Langacker, R. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Vol.1. Stanford: Stanford University Press.

Langacker, R. 2000. *Grammar and Conceptualization*. Berlin: Mouton de Gruyter.

Nam, C., S. Mukherjee, M. Schilk & J. Mukherjee. 2013. Statistical analysis of varieties of English. *Journal of the Royal Statistical Society*, 176: 777–93.

Noonan, M. 1985. Complementation. In T. Shopen ed. *Language Typology and Syntactic Description.* Vol. 2. *Complex constructions*. Cambridge: Cambridge University Press, 42-110.

Schilk, M., J. Mukherjee, C. Nam & S. Mukherjee. 2013. Complementation of ditransitive verbs in South Asian Englishes: a multifactorial analysis. *Corpus Linguistics and Linguistic Theory*, 9, 2: 187-225.

Schilk, M., J. Mukherjee, C. Nam & S. Mukherjee. 2013. Complementation of ditransitive verbs in South Asian Englishes: a multifactorial analysis. *Corpus Linguistics and Linguistic Theory*, 9, 2: 187-225.

Between lexis and discourse: A cross-register study of connectives of contrast

Maïté Dupont (Université catholique de Louvain)

Most reference grammars of English (e.g. Quirk et al. 1972; Quirk et al. 1985; Leech & Svartvik 1994; Halliday & Matthiessen 2004) tend to make very general statements about adverbial connective placement in English, paying relatively little attention to variables such as lexis or register. For example, in Quirk et al. (1972), we find the claim that "the normal position for most conjuncts is I [initial]. […] M [medial] positions are rare for most conjuncts, and E [end] rarer still" (*ibid.*: 526-7). Similarly but in a systemic-functional framework, Halliday & Matthiessen describe adverbial connectives as "what we might call characteristically thematic […]. They are natural Themes" (2004: 83). Yet, as Biber et al.'s (1999: 890-2) brief corpus-based description of adverbial connective placement suggests, connective placement may vary according to both lexical and stylistic factors. The present study investigates 10 adverbial connectives of contrast across three language registers, with a view to assessing the impact of lexis and register on their placement and discourse functions.

The corpus is made up of three subcomponents: c. 1.4 million words from Europarl, a corpus of transcribed debates from the European parliament[1] ; the English subpart of the Mult-ed[2] corpus of quality paper editorials (c. 1 million words); and the English component of the KIAP corpus, made up of research articles from three different disciplines (c. 1.3 million words; see Fløttum et al. 2006). The corpus search focuses on the most frequent connectives from a list of 28 adverbial connectives, after extraction from the corpus via WordSmith Tools 6 (Scott 2012), and manual disambiguation in context. The study is grounded in the framework of Systemic Functional Linguistics (SFL), and relies on a classification of position which identifies three rhematic positions in addition to the usual thematic positions identified in SFL (see Halliday & Matthiessen 2004), thus making it possible to provide a detailed account of the placement of connectives occurring after the topical theme (see Dupont, in press).

Preliminary results reveal that both the frequency and positioning of the connectives investigated vary significantly across the three registers. Parliamentary debates were found to favour thematic positions for adverbial connectives of contrast, as in (1), as opposed to the editorials, which displayed a marked tendency to use connectives rhematically, as in (2). The research articles were found to stand in between these two tendencies.

(1) **However**, while short-term food aid is vital to respond to emergencies (…), EU food aid policy must work towards long-term security in food supply (Europarl).

(2) Tony Blair, **however**, has been persuaded by the Home Secretary that identity cards might be the answer to the Government's own identity crisis (Mult-ed).

[1] I used Cartoni & Meyer's (2012) version of the corpus, which distinguishes clearly between original and translated texts.
[2] http://www.uclouvain.be/en-cecl-multed.html

A focus on the placement patterns of each individual connective revealed that connectives also seem to display fairly idiosyncratic placement patterns. The results highlighted two main types of placement profiles: while some connectives, such as *instead* or *though*, displayed very stable placement profiles across registers, other connectives, such as *on the other hand* and *however*, exhibited fairly variable patterns. The study thus provides evidence of both item- and register-related variation. A more qualitative analysis of the results revealed that connective placement frequently goes hand in hand with specific discourse effects, pertaining to information structure. More particularly, connectives in rhematic positions were found to fulfil discourse functions, such as focusing attention on the theme or some element within the rheme and partitioning of given and new information, in addition to their purely connective function (see also Altenberg 2006; Lenker 2011).

References

Altenberg, B. 2006. The function of adverbial connectors in second initial position in English and Swedish. In K. Aijmer & A.-M. Simon-Vandenbergen eds. *Pragmatic Markers in Contrast.* Oxford: Elsevier, 11-37.

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

Cartoni, B. & T. Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. *Proceedings of the eighth international conference of Language and Resources and Evaluation (LREC)*, Istanbul, May 2012.

Dupont, M.. In press. Word order in English and French. The position of English and French adverbial connectors of contrast. *English Text Construction* 8(1).

Fløttum, K., T. Dahl & T. Kinn. 2006. *Academic Voices. Across Languages and Disciplines.* Amsterdam: Benjamins.

Halliday, M.A.K. & C. M .I. M Matthiessen. 2004. *An Introduction to Functional Grammar*. London: Hodder Arnold.

Leech, G. & J. Svartvik. 1994. *A Communicative Grammar of English*. London: Longman.

Lenker, U. 2011. A focus on adverbial connectors: Connecting, partitioning and focusing attention in the history of English. In A. Meurmann-Solin & U. Lenker eds. *Connectives in Synchrony and Diachrony in European Languages*. Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/08/lenker/ (Accessed 11/07/14).

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

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

Scott, M. 2012. *WordSmith Tools 6*. Liverpool: Lexical Analysis Software.

Location *at* seen through its Swedish and Norwegian equivalents

Thomas Egan (Hedmark University College)
Gudrun Rawoens (Ghent University)

In this paper we investigate whether Swedish and Norwegian translation equivalents of the English preposition *at* can aid us in mapping the semantic field of the preposition when it is used to code physical location. According to Herskovits (1986: 127) *at* is one of three basic topological prepositions in English (the other two are *in* and *on*). She states that: "The use

types of *at* center around one ideal meaning: *at*: for a point to coincide with another" (Herskovits 1986: 128). There are many possible types of spatial coincidence, i.e. many possible locations of the trajector (figure) vis-à-vis the landmark (ground), as pointed out by Lindstromberg (Lindstromberg 2010: 173). We chart the relative frequencies of these various types as they are reflected in different translations into our two target languages.

The corpus data for our study comprise all 571 tokens of *at* coding physical location in the English language original fiction texts found in both The English-Swedish Parallel Corpus (ESPC) and The English-Norwegian Parallel Corpus (ENPC). While there are close equivalents in both Swedish and Norwegian to the other two main topographical prepositions mentioned by Herskovits, *in* and *on*, this is not the case for *at*. In fact we often find tokens of locational *at* translated by the Swedish and Norwegian equivalents of the other two basic prepositions, as in (1a), which is translated into Swedish in (1b) by *på,* the equivalent of *on*, and into Norwegian in (1c) by *i*, the equivalent of *in.*

(1) a. She may be at the nursing station. (AH1)
    b. Hon är kanske på expeditionen. (AH1TS)
    c. Kanskje hun er i vaktstuen. (AH1TN)

Viberg (1998, 2002, 2003) demonstrates that translations can provide evidence of the internal structure of the polysemous network of a single lexeme. Moreover, different senses of a lexeme in a source language which are usually translated into a target language by one and the same lexeme (or construction) may be hypothesised, according to Garretson (2004), to be more closely related within the semantic network of the lexeme in the source language than those translated by different lexemes. Whether one considers English *at* to be polysemous or merely underspecified as to the exact location of the trajector, we will investigate whether the distribution of its translation equivalents, and the degree of overlap between these in the two target languages, can aid us in mapping the semantic field of the English source preposition.

References

Garretson, G. 2004. *The Meanings of English of: Uncovering Semantic Distinctions via a Translation Corpus*. Unpublished MA thesis, Boston University.

Herskovits, A. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge: Cambridge University Press

Lindstromberg, S. 2010. *English Prepositions Explained: revised edition*. Amsterdam: Benjamins.

Viberg, Å.1998. Contrasts in polysemy and differentiation: running and putting in English and Swedish. In S. Johansson & S. Oksefjell eds. *Corpora and Cross-linguistic Research*, Amsterdam: Rodopi, 343-76.

Viberg, Å. 2002. Polysemy and disambiguation cues across languages. The case of Swedish få and English get. In B. Altenberg & S. Granger eds. *Lexis in contrast*, 119-50. Amsterdam: Benjamins.

Viberg, Å. 2003. The polysemy of the Swedish verb komma 'come': a view from translation corpora. In K. M. Jaszczolt & K. Turner eds. *Meaning through Language Contrast*. Vol. 2. Amsterdam: Benjamins, 75-105.

The *-ish*-factor: A corpus-based analysis of *-ish* derivatives in English

Matthias Eitelmann (Johannes Gutenberg University Mainz)
Dagmar Haumann (University of Bergen)
Kari E. Haugland (University of Bergen)

In the course of English language history, the derivational suffix *-ish* has extended its range of application considerably (for previous accounts, see Marchand 1969, Malkiel 1977, Quirk et al. 1985:1553). Used in Old English chiefly to derive ethnic adjectives (e.g. *englisc* 'English', *denisc* 'Danish') and marginally also denominal adjectives (e.g. cildisc 'childish', *deuelisc* 'devilish'), *-ish* became more common as of Middle English and subsequently underwent a number of changes. Thus, *-ish* gradually extended the set of bases to which it attached, starting with primarily monosyllabic adjectives (e.g. *bluish, darkish, oldish*) in the 14th century, with numerals (1) and proper names (2) from the 19th century, and, more recently, with increasingly more complex bases, e.g. compounds (3), phrases (4) or clauses (5):

(1) I guess he was sort of *12**ish*** when he invented the Scuba Scope. (1998 [COCA])

(2) She offered a very *Han Solo**-ish*** grin in response. (2009 [COCA])

(3) But I had to climb the *fire-escape**-ish*** stairway. (1995 [COCA])

(4) It's the soccer uniform: it emphasizes his extremely *young boy**ish*** looks - scrawny legs, underdeveloped arms. (2003 [COCA])

(5) *I know the answer**-ish*** but I need backup (2009 [Web])

Coinciding with the extension of the word formation pattern are semantic changes, which ultimately distort a former more or less consistent form–function mapping: the earliest *-ish*-formations denoted a quality belonging or pertaining to the nominal element – a function eventually ousted by rivalling affixes such as *-y* and *-ous* (cf. Dalton-Puffer 1996:173), whereas the semantics of *-ish* later on rather came to indicate a relation of similitude or, as in the case of adjectival bases, an approximative quality. This strongly subjective sense further develops into a stance marker, a usage that is particularly evident in instances where *-ish* occurs as an autonomous element as in (6):

(6) You must try to remember that some people are normal. ***Ish***. (1990 [BNC])

In a large-scale corpus-based study drawing on a wide array of historical and contemporary corpora (Dictionary of Old English Corpus, Penn Corpora of Historical English, Lampeter Corpus, Early English Prose Fiction, Eighteenth Century Fiction, Nineteenth Century Fiction; BNC, COCA), this paper provides one of the first empirical analyses of the intricately related semantic and functional changes that -ish underwent in its development from a rather neutral affix to a subjective discourse marker. By investigating the distribution of *-ish*-formations from both a morphological and syntactic perspective, this paper sheds light on the productivity of the suffix, which does not only become evident in the numerous hapax legomena (cf. Plag 2008:545), but also in the trajectory of change itself in which *-ish* occurs in new syntactic contexts and new functions. Thus the paper also seeks to provide an

empirical answer to the question of what kind of language change process affects -ish: degrammaticalization, as claimed by Norde 2009, or rather a less clear-cut unidirectional language change as suggested by the notion of constructional change (cf. Traugott/Trousdale 2013, Hilpert 2013).

References

Dalton-Puffer, C. 1996. *The French Influence on Middle English Morphology. A Corpus-Based Study of Derivation*. Berlin/New York: Mouton de Gruyter.

Hilpert, M. 2013. *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press.

Malkiel, Y. 1977. Why ap-ish but worm-y?. In P. J. Hopper ed. *Studies in Descriptive and Historical Linguistics. Festschrift for Winfred P. Lehmann*. Amsterdam: Benjamins, 341-64.

Marchand, H. 1969. *The Categories and Types of Present-Day English Word-Formation. A Synchronic-Diachronic Approach*. 2nd ed. München: Beck.

Norde, M. 2009. *Degrammaticalization*. Oxford: Oxford University Press.

Plag, I. 2008. Productivity. In B.Aarts & A. McMahon eds. *The Handbook of English Linguistics*. Oxford: Blackwell, 537–556.

Plag, I. C. Dalton-Puffer & H. Baayen. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3(2): 209–228.

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

Traugott, E. Closs & G. Trousdale 2013. *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.

Lexical institutionalization reconsidered: *GUI, cyborg, cred, pay-per-view, cyber-* and *techno-*

Roswitha Fischer (Regensburg University, Germany)

At the end of the 20th century, the study of neologisms and their institutionalization was made possible through large electronic text corpora for the first time. Fischer (1998) traced the way of English creative neologisms into the common vocabulary of a speech community. The study broke new grounds of research by considering large stores of data and applying innovative methods of analysis through digital text storage. In the early 1990s large electronic text data had started to become widely available, for instance through the yearly publication of national daily newspapers on CD-ROM. Because earlier corpora like the Brown or the LOB corpus were far too small to investigate new words, these new large text corpora facilitated a thorough and comprehensive study of the institutionalization of neologisms.

After having examined a wide range of novel formations in the London *Guardian* from 1990 until 1996, Fischer (1998) established a model of the institutionalization process. She took the following five main factors into account: (1) existing synonyms and other alternative forms, (2) meaning cues given in the respective texts, (3) frequency and range of occurrence as a reflection of topicality, (4) motivation and transparency, and (5) productivity (Fischer

1998: 171-82). Acronyms, blends and clippings as well as lexical phrases were discussed and analyzed in particular.

Since then, a lot of research has been carried out. At present, two questions arise: (1) what can linguistic studies of neologisms undertaken in the meantime contribute to the results and conclusions in Fischer (1998) in particular?, and (2) what can an examination of the lexical items beyond the time span of 1990-1996 further conduce to an understanding of the institutionalization process?

Regarding the first question, a lot has happened in the scholarly landscape from 1990 onwards. First, we will present relevant research in cognitive linguistics, because the cognitive viewpoint was not considered in Fischer's study at all. Furthermore, we will discuss various corpus-based studies on neologisms which have been published since then. With regard to the second question, we selected a range of lexical items also investigated in Fischer (1998) and searched for them in the London *Guardian* in its electronic version from the 1980s until 2012. In addition, and in analogy to Fischer (1998), we also measured and compared the productivity of the combining forms *techno-* and *cyber-*, this time for the years 1991-2001 and 2002-2012.

On the basis of the new findings, a re-evaluation of the institutionalization process is in order. The institutionalization process emerges as a complex process, in which socio-pragmatic, cognitive and structural factors are closely entwined. It will be shown that the main factor involved in lexical institutionalization is topicality. The second main factor is transparency, which is closely related to form-meaning isomorphy. Less important factors are the existence of possible alternative forms and productivity, which may tip the balance towards or against institutionalization, if topicality and transparency are not so influential.

References

Fischer, R. 1998. *Lexical Change in Present-day English. A Corpus-based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms.* Tübingen: Narr.

Renouf, A. 2013. A finer definition of neology in English: the life-cycle of a word. In H. Hasselgård, J. Ebeling & S. Oksefjell Ebeling eds. *Corpus Perspectives on Patterns of Lexis*. Amsterdam: Benjamins, 177-207.

Schmid, H.-J. 2008. New words in the mind: concept-formation and entrenchment of neologisms. *Anglia* 126: 1-36.

Systematicity beyond obligatoriness in the history of the English progressive

Susanne Flach (Freie Universität Berlin)
Berit Johannsen (Freie Universität Berlin)

Obligatoriness, or obligatorification, a subtype of one of Lehmann's parameters of grammaticalization (Lehmann 2002:124), is often implicitly and explicitly employed as a measure of the status of constructions on grammaticalization clines. The notion is especially

prominent in diachronic studies on tense and aspect, for instance in studies on the progressive. Here, the assumption that the progressive or its use became "obligatory", "systematic" or "grammatically-required" at some time between 1600 and 1900 is at least implicitly widely agreed upon (Strang 1982; Nehls 1988; Elsness 1994; Smitterberg 2005; Kranich 2010).

However, the assumption of obligatoriness poses problems. Besides the issue of how to measure obligatoriness, which is usually deduced from either frequencies or progressive/simple form ratios (Arnaud 1983; Arnaud 1998; Strang 1982; Elsness 1994), there is an inherent bias of comparing language use in earlier stages to seemingly regular usage in Present-Day English (PDE). There seems to be a tendency to attach more weight to singularly exceptional examples found in earlier stages than to those that can also be found in PDE. For instance, 'What do you read, my Lord?' from Shakespeare's Hamlet is often quoted as an example that would require the progressive in PDE (Strang 1982:429–430; Nehls 1988:182; Elsness 1994:5; Rissanen 2000:216; Hundt 2004:47). On a closer look, many of the earlier usage types that are judged to be unsystematic or "deviant from present-day usage" (Elsness 1994:19) can also be found in synchronic corpora. Still, 'exceptional' usage in PDE does not seem to undermine the idea in historical linguistics that PDE usage of the progressive and simple form is in some way (more) systematic than earlier uses.

In contrast, we call into question both the applicability and the usefulness of the notion of obligatoriness by arguing that the use of the progressive is far more systematic in earlier stages of English. Differences in usage patterns are slighter than often assumed; and where they occur, they can be attributed to changes in either individual (lexical) items or changes in constructional semantics. To this end, this talk will take a corpus-driven, collostructional approach to Early and Late Modern English data (Stefanowitsch & Gries 2003; Hilpert 2006; Hilpert 2011; Gries & Hilpert 2012), by looking at changes in association patterns between lexis and constructions. These patterns of course change (and so does the construction), but the results does not readily lend support to the non-obligatory vs. obligatory dichotomy. Rather, we will argue that there is much more to be gained in terms of identifying explanatory factors by taking a bottom-up approach and avoiding the notion of obligatoriness altogether.

Therefore, our paper has two purposes: first, we problematize the concepts of obligatoriness and obligatorification, and second, we use a bottom-up approach to offer different insights on the diachrony of the progressive, which may also contribute to the diachronic study of lexis and grammar on a more general level.

References

Arnaud, R. 1983. On the progress of the progressive in the private correspondence of famous British people (1800–1880). In S. Jacobson ed. *Papers from the Second Scandinavian Symposium on Syntactic Variation, Stockholm, May 15–16, 1982*. Stockholm: Almqvist & Wiksell, 83-94.

Arnaud, R. 1998. The development of the progressive in 19th century English: a quantitative survey. *Language Variation and Change* 10(2): 123-52.

Elsness, J. 1994. On the progression of the progressive in early Modern English. *ICAME Journal* 18: 5-25.
Gries, S. Th. & M. Hilpert. 2012. Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics. In T. Nevalainen & E. Closs Traugott eds. *The Oxford Handbook of the History of English*, 134-44. Oxford: Oxford University Press.
Hilpert, M. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2): 243-56. doi:10.1515/CLLT.2006.012.
Hilpert, M. 2011. 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*, 133-60. Berlin: De Gruyter Mouton.
Hundt, M. 2004. Animacy, agentivity, and the spread of the progressive in Modern English. *English Language and Linguistics* 8(1): 47-69.
Kranich, S. 2010. *The Progressive in Modern English: A Corpus-Based study of Grammaticalization and Related Changes*. Amsterdam: Rodopi.
Lehmann, C. 2002. *Thoughts on Grammaticalization*. 2nd revised ed. (Arbeitspapiere des Seminars für Sprachwissenschaft der Universität Erfurt, ASSidUE 9). Erfurt: Seminar für Sprachwissenschaft der Universität.
Nehls, D. 1988. On the development of the grammatical category of verbal aspect in English. In J. Klegraf & D. Nehls eds. *Essays on the English Language and Applied Linguistics on the Occasion of Gerhard Nickel's 60th birthday*. Heidelberg: Groos, 173-98.
Rissanen, M. 2000. Syntax. In R. Lass ed. *The Cambridge History of the English Language*, vol. 3: 1476-1776. Cambridge: Cambridge University Press, 187-331:
Smitterberg, E. 2005. *The Progressive in 19th-century English: A Process of Integration*. Amsterdam: Rodopi.
Stefanowitsch, A. & S. Th. Gries. 2003. Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209-43. doi:10.1075/ijcl.8.2.03ste.
Strang, B. 1982. Some aspects of the history of the BE+ING construction. In J. M. Anderson ed. *Language Form and Linguistic Variation: Papers Dedicated to Angus McIntosh*. Amsterdam: Benjamins, 427-74.

The lexicogrammar of *BE interested*: Complementation patterns and pedagogical implications

Costas Gabrielatos (Edge Hill University, UK)

BE interested exhibits a variety of complementation patterns (Quirk et al., 1985: 1061, 1063):

- *BE interested in* + Noun Phrase
- *BE interested in* + *–ing* participle Clause
- *BE interested in* + Noun (wh-) Clause
- *BE interested* + *to*-infinitive Clause
- *BE interested* (no complementation)

However, only the first two patterns are included in pedagogical grammars. Pattern (a) tends to be treated at elementary level (e.g. Murphy, 2007), and again at intermediate level, with the addition of pattern (b) (e.g. Murphy, 2012). Although it would be reasonably expected that the remaining patterns would be presented at higher levels, this is not the case (e.g. Hewing, 2013). It is also pertinent to mention that none of the patterns is examined in corpus-

based pedagogical grammars (e.g. Carter & McCarthy, 2006), and no relevant frequency information is provided in descriptive grammars (e.g. Biber et al., 1999).

As the frequency of grammatical constructions can, and should, inform decisions on their inclusion in pedagogical materials (e.g. Leech, 2011), it seems useful to examine the frequency of the complementation patterns of *BE interested* in native speaker corpora, and establish whether their comparative frequency supports the inclusion of some and the exclusion of others. It would also be useful to compare the relative frequency of each pattern in written and spoken language. Finally, pedagogical decisions need to be informed by the extent to which the inclusion/exclusion of particular patterns influences their frequency of use in learner language.

The study used the following corpora: BNC (written and spoken), ICLE (Granger et al., 2009) and LINDSEI (Gilquin et al., 2010). The BNC was accessed via BNCweb (Hoffmann et al., 2008), ICLE via CQPweb (Hardie, 2012), and LINDSEI via AntConc (2014). In each case, random concordance samples of 250 instances of the word *interested* were derived, and then manually cleaned and annotated. For each corpus, the proportion of each complementation pattern was then calculated.

The analysis of the complementation patterns in the BNC showed the following:

- In both written and spoken sub-corpora, '*BE interested in* + Noun Phrase' has the highest proportion (52% and 42%, respectively).
- In the spoken BNC, '*BE interested* (no complementation)' has the second highest proportion (29%).
- The three patterns not treated in pedagogical grammars account for almost half of the instances in spoken English, and almost a quarter in written English.
- In written English, the proportion of '*BE interested* in + *–ing* clause' is almost twice the one in spoken English.
- In spoken English, the proportions of '*BE interested* (no complementation)', '*BE interested in* + Noun (*wh*-) Clause' and '*BE interested* + *to*-infinitive Clause' are almost twice the ones in written English.

Comparisons between the written/spoken BNC and ICLE/LINDSEI suggest that the (absence of) treatment of particular patterns in pedagogical materials influences the proportions used by learners. Both ICLE and LINDSEI show a 50% higher proportion of '*BE interested in* + Noun Phrase' (prominent in pedagogical grammars), and a five times lower proportion of '*BE interested* (no complementation)' (absent from pedagogical grammars). In addition, ICLE has less than half of the proportion of '*BE interested in + –ing* Clause', and LINDSEI has an almost four times lower proportion of '*BE interested* + *to*-infinitive Clause'.

Furthermore, a collocation analysis of the words found in the complements of '*BE interested in + –ing* Clause' and '*BE interested + to*-infinitive Clause' demonstrated that the former shows no preference to particular words, whereas the latter shows very strong preferences, as the top 10 collocates account for 85% of all collocates.

The findings suggest that pedagogical materials could usefully provide more comprehensive and nuanced information on the complementation patterns of *BE interested*.

References

Anthony, L. 2014. AntConc (Version 3.4.3). Tokyo, Japan: Waseda University. http://www.antlab.sci.waseda.ac.jp.

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

Carter, R. & M. McCarthy. 2006. *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge: Cambridge University Press.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot. 2009. *International Corpus of Learner English.* Louvain-la-Neuve: Presses universitaires de Louvain.

Hardie, A. 2012. CQPweb: combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17: 380-409.

Hewing, M. 2013. *Advanced Grammar in Use* (3rd ed.). Cambridge: Cambridge University Press.

Hoffmann, S., S. Evert, N. Smith, D. Lee & Y. Berglund-Prytz. 2008. *Corpus Linguistics with BNCweb. A Practical Guide.* Frankfurt am Main: Peter Lang.

Leech, G. 2011. Frequency, corpora and language learning. In F. Meunier, S. De Cock, G. Gilquin, G. & M. Paquot eds. *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam: Benjamins, 7-31.

Murphy, R. 2007. *Essential Grammar in Use* (3rd ed.). Cambridge: Cambridge University Press.

Murphy, R. 2012. *English Grammar in Use* (4th ed.). Cambridge: Cambridge University Press.

Gilquin, G., S. De Cock, & S. Granger eds. 2010. *LINDSEI: Louvain International Database of Spoken English Interlanguage*. Louvain, Belgium: UCL Presses.

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

Derivational annotation of the BNC: Why and how

Gregory Garretson (Uppsala University)

Modern corpus studies benefit greatly from part-of-speech tagging, syntactic parsing, and newer forms of annotation, such as semantic tagging (e.g. Rayson et al. 2004). Another potentially useful type of annotation which is rarely undertaken is *derivational annotation*, or the linking of derivationally related words. This talk describes a project to derivationally tag the British National Corpus (Burnard 2007), presenting the motivation, the challenges involved, and the procedure used.

This annotation is part of a larger project aiming to find new ways of conceptualizing and investigating collocation. Specifically, the recently introduced concept of *scatter collocation* is an approach to collocation that takes into account the relatedness of words (Garretson, submitted). The idea is that if e.g. *nation* and *lead* are collocates, and *leader* and *national* are collocates too, it is unlikely that these two facts are unrelated in the mental lexicon. Further motivation comes from the comparatively simple morphology of English, such that the collocations *hard work* and *work hard* represent four parts of speech between them, even though the surface forms are identical. This fact is traditionally ignored in studies of collocation, even though this will inflate the apparent collocational strength of *hard* and *work*

in comparison to forms that happen to differ when they change part of speech, such as *quick decision* and *decide quickly*.

The approach taken is to group words into *scatters* (a term taken from Firth 1957) of derivationally related words, such as *nation-national-nationally-etc.* and then to examine the co-occurrence of different scatters in a corpus. A study of 100 high-frequency scatters (Garretson, submitted) showed that scatters co-occur on average via 13 different pairs of lemmas; thus, scatter collocation appears to be a pervasive phenomenon. However, what is required is to test this on a larger scale, including a large number of lower-frequency scatters. This, in turn, requires the development of a derivationally annotated corpus.

In this paper, I will discuss the process undertaken to add derivational tags to the BNC and the questions arising during this process. The first step was to extract information from the hand-crafted lexical database WordNet (Fellbaum 1998), encompassing over 150,000 nouns, verbs, adjectives, and adverbs, grouped into sets of synonyms. Crucially, besides pointers for lexical relations, WordNet includes derivational pointers linking related words. This information was extracted from the WordNet data files by a Perl program written for the task, yielding several thousand sets such as *clear-clearly-clarity-clearness-clarify-clarifying-clarification*. While useful, these sets are smaller than scatters are meant to be. For example, there is no link between the aforementioned set and *unclear-unclearly-unclearness*. Thus, the second step was to use a combination of automated heuristic-based matching and manual work to join these sets into scatters.

Of course, derivational annotation presents interesting problems. For example, it is challenging to find principled ways of deciding (a) how distantly words can be related and still merit inclusion in the same scatter (e.g. *light*, *enlighten*, *unenlightened*), and (b) how to approach polysemy and homonymy (e.g. *light* meaning 'not dark' vs. *light* meaning 'not heavy'). An example of a practice that proved helpful is taking into account the syntactic category of words, such that e.g. the verb *close* was not grouped with the adjective *close*.

To annotate the corpus, scatter tags were added as XML attributes to lexical words. To make the tagging maximally useful, it was decided to add two types of tags: both scatter tags and more conservative derivational tags, so that searches may be performed on "loose" or "tight" groupings of words. The annotations will be made available for public use as provided by the BNC license. It is hoped that this added information will benefit studies of collocation, lexical bundles, and phraseology, which will be able to more easily detect dimensions of variability that had previously gone unnoticed.

References

Garretson, G. [Submitted.] Scatter collocation: exploring relations between lexical families. *Proceedings of ICAME 35*.

Burnard, L. 2007. *Reference Guide for the British National Corpus (XML Edition*). http://www.natcorp.ox.ac.uk/docs/URG/. Accessed 2014-12-14.

Fellbaum, C. 1998. *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Firth, J. R. 1957. *Papers in Linguistics, 1934-1951*. London: Oxford University Press.

Rayson, P., D. Archer, S. L. Piao & T. McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the Workshop on Beyond Named Entity Recognition: Semantic Labelling for NLP Tasks.*In association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25 May 2004, Lisbon, Portugal Paris: European Language Resources Association,. 7-12.

On the focusing adverbs *purely* and *merely* and their adjectival counterparts

Lobke Ghesquière (Research Fund Flanders / KU Leuven / Vrije Universiteit Brussel)

Most traditional grammars of English recognize focusing particles or focusing adverbs such as *only*, *even* and *in particular* as constituting a functional category in its own right. Such focus markers relate the value of the focused expression to a set of alternatives, implicitly or explicitly available in the discourse context. As such they serve a textual function, aiding the discourse-organization by setting up contrastive relations, typically countering expectations and presuppositions in the discourse context.

Focusing adverbs have received considerable attention in the literature, e.g. König 1991, Nevalainen 1991, Sudhoff 2010, Traugott 2006. Some of the widely recognized focusing adverbs, however, have adjectival counterparts which appear to fulfil similar functions in the discourse. The focusing function of the adverbs *purely* and *merely* is, for instance, paralleled by similar uses of the corresponding adjectival forms *pure* and *mere*. Both the adjectives and adverbs can set up contrastive relations, expressing exclusive ('only') (1), inclusive ('even') (2) or restrictive ('just') (3) focus (König 1991, Nevalainen 1991).

(1) a. how hard it proved to cram 12 whole quatrains into **a <u>mere</u> four hours**. (WB)
    b. Does this prove Freud was right? No. **<u>Merely</u> that he is often misinterpreted**. (WB)

(2) a. the excitement I would feel at **the <u>mere</u> anticipation of a visit to the old Empire Theatre**. (WB)
    b. The new rules of air fighting were being made up with each clash. To succeed, **<u>merely</u> to survive**, required an adaptability that was found chiefly in the young. (WB)

(3) a. it was just **<u>pure</u> good fortune** that you ever loved the person who was right for you. (CB)
    b. I subscribe to Sky in London and in Lakeland, where I am at present, **<u>purely</u> to watch the football**. (WB)

Despite the functional overlap between *pure/mere* and *purely/merely* illustrated in the examples above, the focusing potential of the adjectival forms has not been generally recognized. Studies on these and related adjectives label uses such as the above as either identifying (Adamson 2000, Bolinger 1967) or intensifying uses (Gehweiler 2011, Nevalainen 1991:259, Quirk et al. 1985, Vandewinkel & Davidse 2008).

For this study, diachronic corpus research will be carried out into the adjectives *pure* and *mere* and the adverbs *purely* and *merely* to lay bare possible genetic links and any specific differences and similarities between the focusing uses of these items. The forms' semantic and structural diversifications and developments will be described and interpreted in the light of grammaticalization and (inter)subjectification theories. The data looked at for this paper are taken from a selection of historical corpora, including the PPCME and PPCEME corpora and the CLMET3.0 corpus. Synchronic data will be extracted from the WordbanksOnline corpus. The data samples will be analyzed both qualitatively and quantitatively, taking into account frequencies, semantic-pragmatic changes and collocational and scopal potential.

References

Adamson, S. 2000. A lovely little example: word order options and category shift in the premodifying string. In O. Fischer, A. Rosenbach & D. Stein eds. *Pathways of Change: Grammaticalization in English*. Amsterdam: Benjamins, 39-66.

Bolinger, D. 1967. Adjectives in English: attribution and predication. *Lingua,* 18: 1-34.

Gehweiler, E. 2011. *The Grammaticalization of Privative Adjectives: Present-day Uses and Diachronic Development*. PhD thesis. Freie Universität Berlin.

König, E. 1991. *The Meaning of Focus Particles: A Comparative Perspective*. London: Routledge.

Nevalainen, T. 1991. *BUT, ONLY, JUST. Focusing Adverbial Change in Modern English 1500–1900.* Helsinki: Société Neophilologique.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A Grammar of Contemporary English*. London: Longman.

Sudhoff, S. 2010. *Focus Particles in German: Syntax, Prosody, and Information Structure.* Amsterdam: Benjamins.

Traugott, E. C. 2006. The semantic development of scalar focus modifiers. In A. van Kemenade & B. Los eds. The Handbook of the History of English. Oxford: Blackwell, 335-59.

Vandewinkel, S. & K. Davidse. 2008. The interlocking paths of development to emphasizer adjective pure. Journal of Historical Pragmatics; 9: 255-87.

Multidimensional analysis for the masses

Andrew Hardie (Lancaster University)

The standard "toolkit" of corpus linguistics, though flexible, is surprisingly narrow: the four key techniques of concordances, frequency lists, statistical collocation, and keywords underlie a high proportion of work in the field. As McEnery and Hardie (2012: 41-46) point out, for the field to advance it is essential that new methods should become available within the context of the software systems that corpus linguists commonly use. Yet this rarely happens; the current generation of web-based concordancers mostly replicate the functionality of earlier software such as Scott's WordSmith Tools or Barlow's MonoConc.

This is not for lack of new corpus-based methods, which have been devised and promulgated regularly. One well-established example is Biber's "Multidimensional Analysis", as set forth in his classic (1988) study and subsequent work. Biber (1988, 1995) demonstrates convincingly the value of this framework in studying variation. Yet it has scarcely been adopted outside the circle of Biber's collaborators. Where other authors adopt the framework

it has often been via simpler methods (e.g. Xiao and McEnery 2005; later work by Xiao hews more closely to Biber's model). Given the widespread and enthusiastic adoption of another of Biber's innovative analyses, Lexical Bundles, the best explanation for the relatively low uptake of multidimensional analysis is its conceptual, statistical and technical difficulty level. To replicate Biber's approach, it is necessary not only to execute complex corpus queries across annotation as well as words, but also to extract frequency breakdowns for those queries across texts, to add further quantitative measures, and to perform and interpret a factor analysis – which requires both some understanding of the statistics of factor analysis, and practical skills with advanced software such as SPSS or R.

This paper will detail an approach taken to embed these procedures seamlessly into an existing web-based concordancer (CQPweb: Hardie 2012), to open up multidimensional analysis to linguists who would otherwise be barred from these methods by lack of technical know-how. The powerful R statistical environment is used as the back-end engine for the factor analysis – but the complexity of R, and of the compilation of data from a corpus as input to R, is hidden behind a user-friendly front-end webpage.

With this implementation complete, a user of the system approaches multidimensional analysis as follows. First, they carry out a series of corpus queries, saving the results of each one. Second, they invoke a novel interface which allows these saved results to be internally translated into quantified features across texts or other corpus segments. As in Biber's model, additional features may be mathematically derived (e.g. from the difference between the frequencies of hits for two queries) or selected from a menu of special feature types (e.g. type-token ratio).

The user then selects the features that they wish to incorporate into a feature matrix for the corpus texts; this matrix is cached in the system database for subsequent analysis. A multidimensional analysis can then be carried out by requesting a factor analysis of the matrix; the user interface affords control over various options, e.g. factor rotation method, number of factors, and so on. Finally, the factor analysis results are presented; to support the functional interpretation essential to establishing dimensions of variation within a multidimensional model, not only the factors and feature weightings are displayed, but also a set of user-friendly visualisations plotted using R.

Finally, by supplementing these advances in software design with a comprehensive listing of the query patterns necessary to replicate Biber's (1988) methodology on any English corpus annotated with CLAWS-style POS tags, the work described here renders multidimensional analysis, for the first time, truly a method for the general body of corpus linguists.

References

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

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

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. and Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
Xiao, R. & McEnery, T. 2005. Two approaches to genre analysis: three genres in modern American English. *Journal of English Linguistics*, 33, 62–82.

Cross-varietal patterns in the English genitive alternation: A pilot study

Benedikt Heller (KU Leuven)

Variation between the *s*-genitive, as in (1a), and the *of*-genitive, as in (1b), is one of the best-researched syntactic alternations today (cf. Rosenbach 2014).

(1) a. And one person's allergic to **my friend's dog**
<ICE-GB:s1a-081>

b. What is **the present state of public health care in India**
<ICE-IND:s1b-041>

The choice between the *s*-genitive and the *of*-genitive is well known to be influenced by a variety of language-internal and language-external factors (e.g. animacy, syntactic weight, givenness, etc.), whose influence has been shown in various studies (e.g. Wolk et al. 2013, Grafmiller 2014, inter alia). However, so far comparatively little large-scale comparative research has been conducted on how these well-known factors differ in strength across different native varieties and second-language varieties of English. This study seeks to explore the linguistic constraints that govern the English genitive alternation across nine different varieties of English by looking at statistical evidence from corpora of the ICE family (cf. Greenbaum 1996). It takes into account a number of factors that have been shown to influence the alternation, for example, syntactic weight of the constituents (i.e. possessor and possessum), discourse accessibility of the possessor, or possessor thematicity.

The cross-varietal differences are analyzed using state-of-the-art statistical methods such as mixed-effect multivariate logistic regression analysis and conditional inference trees. These methodological approaches allow visualization and comparison of the effect sizes with which the aforementioned factors influence the choice between the two genitive variants. The study compares effects of various factors across four native varieties of English (i.e. British English, Canadian English, Irish English, and New Zealand English) and five second-language varieties (i.e. Hong Kong English, Indian English, Jamaican English, Philippine English, and Singapore English).

The results of the study are interpreted within the probabilistic grammar framework, which seeks to explore the hidden, but cognitively real probabilistic constraints that language users are implicitly aware of when choosing grammatical variants. It has been shown in experimental studies (cf., e.g., Bresnan & Ford 2010) that language users' intuitions match predictions from statistical models, so the genitive variation models that this paper will report can be viewed as a generalized numerical representation of linguistic expertise in a speech community. The ultimate goal of this study is, therefore, to shed light on differences with

regard to linguistic knowledge between speakers of varieties of English and compare its results to predictions from the well-established Dynamic Model of the evolution of post-colonial Englishes (Schneider 2007).

References

Bresnan, J.& M. Ford. 2010. Predicting syntax: processing dative constructions in American and Australian varieties of English. *Language*, 86 (1): 186-213.

Grafmiller, J. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics*, 18: 471-96.

Greenbaum, S. 1996. Introducing ICE. Comparing English Worldwide: *The International Corpus of English*: 3-12.

Rosenbach, A. 2014. English genitive variation: the state of the art. *English Language and Linguistics*, (18): 215-62.

Schneider, E. 2007. *Postcolonial English*. Cambridge: Cambridge University Press.

Wolk, C., J. Bresnan, A. Rosenbach & B. Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English. *Diachronica*, 30, 3: 382-419.

Constructions and the Dynamic Model: Comparative Correlative Constructions in Englishes around the world

Thomas Hoffmann (KU Eichstätt-Ingolstadt)

Schneider's influential Dynamic Model (2007) of the evolution of post-colonial varieties of English predicts that it is the lexicon-syntax interface that exhibits first traces of the emergence of characteristic structural innovations during the phase of nativization. Independently of this, Construction Grammar approaches (for an overview, cf. Hoffmann and Trousdale 2013) have pointed out that mental grammatical knowledge does not comprise an independent set of lexical items and syntactic rules. Instead, the basic building block of mental grammars are form-meaning pairings ('constructions') that range from items whose phonological form is completely filled (e.g. the word *apple* [æpl]—'apple') to ones that are completely schematic, i.e. whose phonological pole consists only of slots that can be filled by various other constructions (e.g. the Resultative construction [SBJ V OBJ RP]—X causes Y to become Z_{STATE}; Goldberg 2006: 73). In between those two extremes, there are a great number of constructions with varying degrees of schematicity (such as the comparative construction [X BE $Adj_{comparative}$ ðən Y]—'X is more Adj than Y').

In this talk, I will discuss how the predictions of the Dynamic Model receive cognitive support by this constructionist notion of the lexicon-syntax cline. I will illustrate this by focussing on a largely schematic filler-gap construction, Comparative Correlative (CC) clauses ([*the* [ ]$_{comparative\ phrase1}$ (clause)]$_{C1}$ [*the* [ ]$_{comparative\ phrase2}$ (clause)]$_{C2}$, e.g. *The higher the price is, the more interesting the product is*.; cf. Sag 2010) that also exhibits strongly lexicalised, idiomatic instances (such as *the more, the merrier*; Fillmore, Kay and O'Connor 1988: 506). The data examined are taken from the International Corpus of English project for L1 British English, Canadian English, Irish English and New Zealand English (stage V varieties in the Dynamic Model), as well as various post-colonial L2 varieties (for stage IV:

Jamaican English and Singapore English / for stage III: Hong Kong English, Indian English, Kenyan English, Nigerian English and Philippine English).

All data are subjected to a quantitative statistical analysis (using Hierarchical Configural Analysis; cf. Gries 2008: 242-54) that detects significant main effects as well as factor interactions within and across varieties.

As I will show cognitive factors such as type and token frequency, prototype and processing effects account for structural similarities across these varieties (including a strong preference of AdjP phrase fillers in both clauses as well as a dispreference of NPs or parallel deletion phenomena across C1 and C2). On top of that, a constructionist view also provides a straightforward account of innovative uses (such as ***as*** *you a mother express the milk,* ***the more*** *milk is made* ICE-KE: br-talkk).

References

Culicover, P. W. & R. Jackendoff. 1999. The view from the periphery: the English comparative correlative. *Linguistic Inquiry*, 30: 543−71.

Fillmore, C., P. Kay & M. C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language,* 64: 501−38.

Goldberg, A. E. 2006. *Constructions at Work: The Nature of Generalisation in Language*. Oxford: Oxford University Press.

Hoffmann, T. and G. Trousdale. 2013. Construction grammar: introduction. In T. Hoffmann & G. Trousdale, eds. 2013. *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 1-12.

Sag, I. A. 2010. English filler-gap constructions. *Language* 86, 3: 486-545.

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

Variation in South Wales Coalfield English: Rhoticity and the realization of the FACE and GOAT vowels

Magnus Huber (University of Giessen)

Of the varieties of English in the British Isles, Welsh English remains among the least-studied, which is partly due to the lack of easily accessible collections of linguistically processed natural speech. The Corpus of English in the South Wales Coalfield, situated in the old county of Glamorgan in South-East Wales, was compiled by the author to alleviate this situation. The present paper uses a corpus linguistic approach to investigate phonological variation in Coalfield English, focussing on three phonological variables: rhoticity as well as the FACE and GOAT vowels.

The South Wales Coalfield had been sparsely populated and almost completely Welsh-speaking until ca. 1850. Industrialization and the growing demand for coal during the 19th century led to massive labour immigration, especially from the nearby counties in Wales and England. In the 50 years from 1861 to 1911, 406,000 immigrants settled in Glamorgan, whose population more than tripled in the same period. About 70% of the immigrants spoke

English as their mother tongue. This influx led to contact between Welsh and English but also to contact between different dialects of English. The 1890s were a period of rapid linguistic transition, the number of monoglot Welsh speakers dropping drastically from 21.7% to 6.6%. Today, the area is almost completely monolingual anglophone.

When the last coal mines were closed in South Wales in the 1970s and 80s and the industrial culture of the area was about to vanish, several oral history projects were carried out to "preserve much of the written, visual and oral history of the South Wales Coalfield" (University of Wales Swansea Library & Information Services information leaflet). The South Wales Coalfield Collection (SWCC) University of Wales Swansea, holds 678 audio recordings from these projects. The Corpus of English in the South Wales Coalfield was compiled from these SWCC recordings, including orthographic transcripts of all interviews with speakers born between 1875 and 1900 (25), and of a sample of those born between 1900 and 1930 (15). These 40 recordings comprise about 57 hours of speech and 500,000 words. The oldest speakers were born in the 1870s, just one generation after large-scale immigration into the Coalfield started, and provide a unique linguistic window into the early stages of the formation of Coalfield English. Together with the very detailed census data available for this period this offers the rare chance of studying the "birth" of the industrial dialect of the SE Wales mining valleys.

With regard to the phonological variables investigated, the analysis of the corpus shows that Coalfield English is variably rhotic with significant gender differences, males showing 17% r-full realizations of post-vocalic r and females only 6%. There is also an interesting split between Coalfield-born speakers (6% rhoticity) and immigrants from the rest of Wales (35%) and England (40%). Concerning the FACE and GOAT vowels, there is even more variability. In most accents in the British Isles, the PANE/PAIN and the NOSE/KNOWS subsets merged in the so-called Long Mid Merger (by ca. 1600) so that the vowels of these pairs are identical today. Some speakers of Coalfield English, however, largely preserve the Middle English opposition between monophthongal PANE/NOSE on the one hand and diphthongal PAIN/KNOWS on the other. Other speakers show varying degrees of merger of PANE and PAIN as well as of NOSE and KNOWS, which are due both to the monophthongization of PAIN/KNOWS and to the diphthongization of PANE/NOSE.

The findings from the corpus study will be contextualized within the larger picture of English in Wales by drawing on dialectological data from the Survey of Anglo-Welsh Dialects (Parry 1977, 1979, 1999).

References

Parry, D. ed. 1977. *The Survey of Anglo-Welsh Dialects*, Vol. 1: The South-East. Swansea.
Parry, D. ed. 1979. *The Survey of Anglo-Welsh Dialects*, Vol. 2: The South-West. Swansea.
Parry, D. ed. 1999. *A Grammar and Glossary of the Conservative Anglo-Welsh Dialects of Rural Wales*. Sheffield.
University of Wales Swansea Library & Information Services. (n.d.). *The South Wales Coalfield Collection*. Internet access to a unique research collection. Swansea. (Information leaflet).

The processing of English collocations by native speakers and learners of English

Jennifer Hughes (Lancaster University; Centre for Corpus Approaches to Social Science)

There is growing evidence to suggest that formulaic sequences are processed more quickly than non-formulaic sequences by native speakers of English (Conklin and Schmitt 2012:56). However, comparatively few studies have investigated the processing of formulaic sequences by learners of English (Wray 2002:144; Schmitt et al. 2004:55). These studies have either reported mixed results (Conklin and Schmitt 2012:45), or demonstrated a processing advantage only for fixed idioms or other highly restricted formulaic sequences (e.g. Conklin and Schmitt 2008:83; Underwood et al. 2004:160-161).

In this paper, I present the results of a self-paced reading experiment which aimed to find out whether formulaic sequences are processed more quickly than non-formulaic sequences by 20 native speakers and 20 learners of English. However, instead of focusing on fixed idioms or other highly restricted formulaic sequences, I assume a broader conceptualisation of formulaic language by focusing on transitional probabilities, i.e. the probability of word Y being produced given that the previous word was X. The research questions are:

(1) Do native speakers of English and learners of English process the nouns in adjective-noun bigrams more quickly in bigrams that have a higher transitional probability compared to bigrams that have a lower transitional probability?

(2) If learners of English are found to be sensitive to the transitional probabilities between words, is this sensitivity related to their English proficiency level and/or to their level of acculturation into the English-speaking community?

(3) If there is a difference in reading time between the nouns in both conditions, is this difference in reading time sustained to the words that follow the noun? (i.e. are there any spillover effects?)

The transitional probabilities were calculated by dividing the number of times the bigram X-then-Y occurs in the written section of the BNC by the number of times X occurs in the written section of the BNC altogether (McEnery and Hardie 2012:195). I extracted 10 adjective-noun bigrams with a higher transitional probability (median = 0.0175) and 10 with a lower transitional probability (median = 0.0009). The adjectives were the same in each condition; the nouns were different but were matched for frequency and length. The bigrams were then embedded into plausible sentences for use in the self-paced reading experiment. In order to answer research question 2, I asked the learners to complete an English proficiency test and an acculturation questionnaire. The questionnaire responses were then converted into an overall acculturation score for each learner.

The results show that the bigrams with the higher transitional probabilities are processed significantly more quickly than the bigrams with the lower transitional probabilities by both the native speakers and the learners. Furthermore, the words following the bigrams with the higher transitional probabilities are often processed more quickly than the same words

following the bigrams with the lower transitional probabilities. There is a significant interaction between the reading times and the proficiency level of the participants. However, no significant relationship was found between the reading times and the learners' level of acculturation into the English-speaking community. In sum, these results therefore provide further confirmation for the psychological reality of formulaicity and collocation and their importance in language learning.

References

Conklin, K. & N. Schmitt. 2008. Formulaic sequences: Are they processed more quickly than non-formulaic language by native and non-native speakers? *Applied Linguistics*, 291: 72-89. doi:10.1093/applin/amm022.

Conklin, K. & N. Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics*, 32: 45-61. doi:10.1017/S0267190512000074.

McEnery, T. & A. Hardie. 2012. *Corpus Linguistics: Methods, Theory and Practice.* Cambridge: Cambridge University Press.

Schmitt, N., Z. Dörnyei, S. Adolphs, & V. Durow. 2004. Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt ed. *Formulaic Sequences: Acquisition, Processing, and Use*. Amsterdam: Benjamins, 55-86.

Underwood, G., N. Schmitt, N. & A. Galpin. 2004. The eyes have it: an eye-movement study into the processing of formulaic sequences. In N. Schmitt ed. *Formulaic Sequences: Acquisition, Processing, and Use*. Amsterdam: Benjamins, 153-72.

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

Topic modelling in a corpus of academic text

Susan Hunston, Akira Murakami, Paul Thompson & Dominik Vajn (University of Birmingham)

One of the challenges in corpus linguistics is to develop 'bottom up' methods of grouping individual texts without imposing distinctions relating to genre, author membership etc. In this talk, we explore the use of a machine-learning technique called 'topic modelling' (Blei, 2012; Griffiths & Steyvers, 2004; Grün & Hornik, 2011) by demonstrating its use in a corpus of academic discourse. Topic modelling automatically identifies "topics" in a given corpus according to patterns of word co-occurrence in individual texts (e.g., If word X is frequent in a text, word Y is also likely to be frequent in the same text). Certain sets of words are found to co-occur frequently and can be considered as keywords of the "topic". They are distinguished from other co-occurring sets, that is, from other "topics". Each text, in turn, is characterized by topic distribution: a text consists of multiple topics of different probability, and a text with the high probability of a given topic can be considered as a key text of that topic.

The present study identifies topics in research papers in environmental science and examines (i) how the various topics might be described or classified, (ii) differences in topic distribution between journals, (iii) within-journal chronological changes in terms of topics, and (iv) topics that are prominent in different parts of a paper. Our corpus consists of research papers published in 11 journals over 10 years. The holdings of these journals have been made

available to us by Elsevier publishers. The corpus consists of 56 million words over 11,700 papers. The method requires an arbitrary choice of the number of topics to be identified: this paper is based on the identification of 100 topics.

Our findings are as follows. Topics are identified that are discipline-related, method-related, 'real-world'-related and discourse-related. The 11 journals do not emerge as discrete entities using this method, though there are a substantial number of journal-specific topics. The journal *Global Environmental Change* (GEC), for instance, includes many papers with high probability in one of the topics with such keywords as *develop*, *global*, *world*, and *industry*. This is because GEC papers often address issues in international development, and the papers on the topic tend to have high frequency of these words.

Topic modeling is informative about the chronological change of topics. We note, for example, that the topic of 'international development', discussed above, decreases consistently in GEC over 10 years, while we observe an increase in the topic of quantitative empirical analysis that includes such keywords as *variable*, *correlate*, *regression*, *significance*, and *linear*. This suggests that the journal has changed its focus away from discussion-oriented papers about international development towards quantitative empirical studies.

Topic modeling also allows us to identify some position-specific topics within papers. For example, a topic with keywords including *pollution*, *atmosphere*, and *air* occurs prominently at the beginning of papers, rather than in other positions. In other words, the topic is used to set the scene of empirical studies and is associated with the literature review. On the other hand, we have identified a topic with keywords including *limitation*, *possible*, and *future* which is prominent at the end of papers, most probably because they reflect the concerns of conclusion sections of papers.

Focusing on 100 topics raises difficulties in terms of offering a large mass of information for interpretation. On the other hand, the word-groups or topics identified by this method support the interpretation of discipline, time and discourse position (or genre) as key to distinctions within a corpus.

References

Blei, D. M. 2012. Probabilistic topic models: surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the ACM*, 55(4): 77-84. doi:10.1145/2133806.2133826.

Griffiths, T. L. & M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (supplementary 1), 5228-35. doi:10.1073/pnas.0307752101.

Grün, B. & K. Hornik. 2011. topicmodels : An R Package for fitting topic models. *Journal of Statistical Software*, 40(13). http://www.jstatsoft.org/v40/i13.

Keyness across business genres: Audit reports and directors' reports

Marlén Izquierdo (University of the Basque Country, EHU-UPV)
Rosa Rabadán (University of León)
Isabel Pizarro Sánchez (University of Valladolid)

Writing reports is a common communicative practice within the business discourse community. In turn, depending on the specific aim such a text pursues, as well as what specific members within the aforementioned community it is addressed to, a varied typology of reports seems to emerge. In this sense, Audit Reports and Directors' Reports represent two prominent, context-bound types of business discourse (Flowerdew and Wan, 2010). While the defining features seem to pin down a clear-cut distinction between one and the other, the two types of report are believed to hold some sort of interrelationship worth delving into, both for LSP research and – specially – for LSP teaching purposes (Swales, 1990).

Within this framework, a contrastive study is undertaken at the lexical level. On the assumption that both genres share a reporting function, but their aboutness is a priori different, this study contrasts the behaviour of their keywords, paying special attention to shared lexical items. The analysis unfolds in three stages; first, using the AntConc software (Anthony, 2010), (shared) keywords are selected from frequency wordlists which will be compared to a reference corpus. In this way, keyness in Audit Reports and in Directors' Reports is calculated (Scott, 2006). In a second stage, the (shared) lexical items are described in context, per genre, examining three aspects of their linguistic behaviour; first, the grammatical nature of a keyword, whether it is a noun or a verb, etc.; second, its collocational and colligational patterns, and whether the keyword displays any tendency for lexical relations or whether it belongs to a firmly fixed lexical relationship; and third, to what extent the keywords under analysis can be treated as specialized, business lexical items. In a third stage, the features observed in the previous description will be contrasted in search of similarities and differences across genres, revealing so-called "key associates" (Scott, 2006). Expected findings shed light on the extent to which keyness is tightly bound by context, so that while a business keyword is actually key in one genre, it might not be so in another. Likewise, the results suggest that genres diverge with regard to their actual load of specialized, fixed lexical combinations (phraseology) (Granger and Meunier 2008).

The article concludes discussing the usefulness and usability of a keyword-based approach to LSP, and, hence, the pedagogical implications of teaching genre-bound, specialized keywords, in business English courses, as a more innovative and tailored education of would-be auditors and/or directors, in this case.

References

Anthony, L. (2010). AntConc (Version 3.2.1.) [Computer Software] Tokyo, Japan: Waseda University. http://www.antlab.sci.waseda.ac.jp/

Bondi, M. & M. Scott eds. 2010. *Keyness in Texts*. Amsterdam: Benjamins.

Flowerdew, J. & A. Wan. 2010. The linguistic and the contextual in applied genre analysis: the case of the company audit report. *English for Specific Purposes; 29*. 78-93.

Gotti, M. 2003. *Specialized Discourse. Linguistic Features and Changing Conventions*. Bern: Peter Lang.
Granger, S. & F. Meunier. 2008. *Phraseology: an Interdisciplinary Perspective*. Amsterdam: Benjamins.
Hyland, K. 2008. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1): 4-21.
Scott, M. 2006. The importance of keywords for LSP. In E. Arnó Macià, A. Soler Cervera & C. Rueda Ramos eds. *Information Technology in Language for Specific Purposes: Issues and Prospects.* NY: Springer.
Swales, J. 1990. *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.

A lost Canadian dialect: The Ottawa Valley 1975-2013

Bridget Jankowski & Sali Tagliamonte (University of Toronto)

The Linguistic Survey of the Ottawa Valley (LSOV), collected in 1975–1981, documents at least ten distinct dialects based on settlement patterns, as Irish, Highland Gaelic, Lowland Scots, German and Kashubian Polish-speaking immigrants came in contact with French and English-speaking populations (Pringle & Padolsky, 1983: 326–29). However, aside from a few publications (e.g. Pringle & Padolsky, 1983), it remains an untapped reservoir of dialect data.

Through arrangements with the original compiler of the LSOV, Professor Emeritus Ian Pringle, we have undertaken the task of digitizing more than 175 reel-to-reel recordings (nearly 200 speakers), and creating transcripts to facilitate quantitative analysis. The possibility of comparing these materials with more recently collected data from the Ottawa Valley (OV) (Tagliamonte 2013–2018) offers an unprecedented opportunity to track variation and change over 40 years in real-time in Canada, a country once thought to have minimal regional differentiation (Dollinger & Clarke, 2012; Chambers, 1991).

In this paper, we examine two widely-studied linguistic features found in many English dialects: preterit *come*, as in (1) and verbal *–s* with 3rd person plural, as in (2).

1. a. They *come* here in nineteen-and-seven. (male, b. 1898)
   b. They *come* and asked us... (female, b. 1961)
2. a. You know, that my eyes *is* failing. (female, b. 1890)
   b. In my time, the barns *was* all built beforehand. (female, b. 1903)
   c. They *was* getting up in years then. (male, b. 1939)

These variants alternate with the standard forms (*came* and *are/were*) in virtually all English dialects, from Britain (Szmrecsanyi, 2013; Tagliamonte, 2001), to Tristan da Cunha (Schreier, 2010) to North America (Atwood, 1953). Most reports are simply attestations or frequency counts. However, constraints on usage can distinguish origins, locality and the nature of linguistic change (see Poplack & Tagliamonte, 2001). Moreover, these particular linguistic features are also ideal for identifying socio-cultural influences, in particular local

affinity (see e.g. Labov, 1963; Schilling-Estes, 1999). Using the tools of comparative sociolinguistics (Tagliamonte, 2012) and statistical modeling techniques (e.g. Tagliamonte & Baayen, 2012), this paper assesses the nature and character of the Ottawa Valley situation.

Both nonstandard variants are present in the LSOV and the modern OV, however there is a decline in frequency for 3p. plural *–s* (13% in the LSOV vs. 4.5% in the OV, N = 1217), yet robust variation and stability for preterit *come* (47% in LSOV and 51% in OV, N = 678). At the same time, both features retain linguistic conditioning that can be traced to the historical record, exposing longitudinal maintenance of the variable grammar. In the case of 3p. plural *–s*, variation is restricted to forms of the verb *be*, as in (2). The comparative perspective of two points in time (1975–81 vs. 2013) further exposes how this variation becomes incrementally restricted to the past tense, giving us a window on the longitudinal ramifications of obsolescence. Further, preterit *come* shifts from having only internal linguistic conditioning in LSOV to both linguistic and social conditioning in the modern OV, demonstrating that features develop social correlates over time under changing socio-cultural conditions (see also Schilling-Estes & Wolfram, 1999).

In summary, even in the face of modern communication and urban sprawl, the Ottawa Valley still has a distinct dialect. While certain features are being lost, others are strongly retained, leading us to argue, consistent with earlier research, that dialects are a critical litmus test for tracking the mechanisms of linguistic change, as well as generational adjustments to external forces. Moreover, this study brings to light a dialect treasure within an erstwhile monolithic country.

References

Atwood, E. B. 1953. *A Survey of Verb Forms in the Eastern United States.* Ann Arbor: University of Michigan Press.

Chambers, J. K. 1991. Canada. In J. Cheshire ed. *English Around the World.* Cambridge: Cambridge University Press, 89-107.

Dollinger, S. & S. Clarke. 2012. On the autonomy and homogeneity of Canadian English. *World Englishes*, 31(4): 448-66.

Labov, W. 1963. The social motivation of a sound change. *Word* 19: 273-309.

Poplack, S. & S. Tagliamonte. 2001. *African-American English in the Diaspora.* Malden & New York: Blackwell.

Pringle, I. & E. Padolsky. 1983. The linguistic survey of the Ottawa Valley. *American Speech,* 54(4): 325-44.

Schilling-Estes, N. 1999. Reshaping economies, reshaping identities: gender-based patterns of language variation in Ocracoke English. Engendering communication. *Proceedings of the Fifth Berkeley Women and Language Conference.* Berkeley, CA: Berkeley Women and Language Group. 509-20.

Schilling-Estes, N. & W. Wolfram. 1999. Alternative models of dialect death: dissipation vs contraction. *Language*, 75(3): 486-521.

Schreier, D. 2010.Tristan da Cunha English. In D. Schreier, P. Trudgill, E. W. Schneider, & J. P. Williams eds. *The Lesser-Known Varieties of English: An Introduction.* Cambridge: Cambridge University Press.

Szmrecsanyi, B. 2013. *Grammatical Variation in British English Dialects.* Cambridge: Cambridge University Press.

Tagliamonte, S. A. & R. H. Baayen. 2012. Models, forests and trees of York English: w*as/were* variation as a case study for statistical practice. *Language Variation and Change,* 24(2): 135-78.
Tagliamonte, S. A. 2001. *Come/came* variation in English dialects. *American Speech,* 76(1): 42-61.
Tagliamonte, S. A. 2012. Comparative sociolinguistics. In J. K. Chambers & N. Schilling-Estes eds. *Handbook of Language Variation and Change,* 2nd ed. Malden and Oxford: Blackwell.
Tagliamonte, S. A. 2013–2018. Social Determinants of Linguistic Systems. Insight Grant: Social Sciences and Humanities Research Council of Canada.

Coping with errors: Estimating the impact of OCR errors on corpus linguistic analysis of historical newspapers

Amelia Joulain-Jay (Lancaster University)

The increasing availability of digitized historical material opens up promising new avenues of research for historical linguists, discourse analysts, literary scholars, and historians; notably, it allows the application of digital methods to historical questions. The first part of the British Library (BL)'s digital collection of c19th newspapers is an example of such a resource. It was made available to the public in early 2008 and contains the full runs of 48 periodicals printed between 1800 and 1900, amounting to around 2 million pages (Conboy 2009). However, like much digitized historical material, it has been produced using optical character recognition (OCR) procedures, which have reportedly low levels of accuracy when used on historical material (e.g. Holley 2009). Tanner et al. (2009, section 6) report an average word accuracy of 78% in the BL's c19th newspaper collection, with the average accuracy of 'significant words' (words judged likely to be of interest to researchers, i.e. not words such as *the*, *he* and *it*) having an even lower accuracy rate of 68.4%.

A growing body of research investigates the question of how to automatically correct OCR errors in order to raise accuracy levels in the post-OCR phase. Although fairly successful methods have been reported, these can be costly to apply and in any case do not achieve perfect results. An example is the commercial software reported by Evershed and Fitch (2014) which claims to "reduce the number of articles missed by a keyword search due to OCR errors by over 50%" (OverProof 2014), but which is prohibitively priced for most non-corporate entities: e.g. it would cost several thousand dollars to use on just one of the newspapers in the British Library's collection.

This paper addresses the related but distinct question of how problematic these errors are in the first place: if we cannot remove these errors, are there ways to work around their presence, and if not, is it possible to assess their likely effect on the types of analyses we wish to carry out? I focus on corpus linguistic procedures, particularly collocation, in order to discuss how OCR errors may affect the results of large-scale corpus-assisted discourse analysis of c19th news reportage.

First, I consider the question of how to identify errors and estimate error rates for specific words. Identifying items which are likely to be errors can be done automatically with acceptable levels of precision and recall; the challenge is retrieving erroneous forms for a

given word of interest. Errors are often at edit-distances of 4, 5 or 6 from the correct form (Evershed and Fitch 2014, section 1) so using edit-distance is not very helpful. Fuzzy queries (using wildcards) are more helpful, but estimating their recall and precision remains a challenge. Next, I examine the distribution of errors in the corpus. I find that error rates vary from word to word, as well as over time, in the corpus. This is potentially problematic since it affects the reliability of frequency counts, which has knock on effects on the statistics of collocation and keyness analysis. I present examples of various corpus linguistic procedures applied to 'The Era', one of the newspapers in the British Library's collection of c19th newspapers, in order to assess the extent to which the results of these procedures are impacted by the error rates identified in the corpus. Finally, I recommend some measures for dealing with OCR-derived data.

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.

Evershed, J. & Fitch, K. 2014. Correcting noisy OCR: context beats confusion, *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage ACM*. 45-51.

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., T. Muñoz, & P. H. Ros. 2009. Measuring mass text digitization quality and usefulness, *D-Lib Magazine*, 15, 7/8. 1082-9873.

Work on -ing versus *work at -ing*: Teasing two constructions apart with corpus data

Mark Kaunisto & Juhani Rudanko (University of Tampere)

Consider sentences (1a-b), both from the Corpus of Contemporary American English, COCA:

(1) a. Hopefully, they work on making the students comfortable with dealing with what they want to do. … (1994. SPOK)

b. After preliminary joins were made, the team worked at deciphering and interpreting the texts. (1994. COCA)

The sentences in (1a-b) are similar in that both involve non-finite *-ing* clauses, or gerunds, introduced by a preposition, the prepositions being *on* in (1a) and *at* in (1b). Further, they are similar in that both involve subject control, with the understood subject of the *-ing* clause being controlled by the matrix subject. However, the prepositions are different, and this means that the constructions are different. Adopting the view that different constructions involve different meanings, in the spirit of both cognitive grammar and construction grammar approaches to linguistic description, it is then of interest to inquire into the question of how

the two constructions may differ with respect to their meanings. This is the main research question of the present paper.

The investigation draws on data from large electronic corpora. COCA provides the point of departure, and it is supplemented with data from selected decades of the Corpus of Historical American English, COHA, to gain a perspective on the history of the two constructions and their frequencies in earlier English. Both constructions involve gerundial clauses, and concepts introduced for instance in Allerton (1988) as potentially relevant to the analysis of the semantics of such clauses are considered, in order to shed light on the nature of the two constructions. Another aspect of the constructions taken into account is the nature of the preposition in the two patterns. The prototypical meanings of the two constructions are considered, drawing on work for instance by Herskovits (1986), with the aim of relating the use of the preposition in each construction to a more prototypical meaning of the preposition in question, and thus opening a perspective on the constructional meaning in each case.

References

Allerton, D. 1988. 'Infinitivitis' in English. In J. Klegraf & D. Nehls eds. *Essays on the English Language and Applied Linguistics on the Occasion of Gerhard Nickel's 60th Birthday*. Heidelberg: Julius Groos, 11-23.

Herskovits, A. 1986. *Language and Spatial Cognition: an Interdisciplinary Study of the Prepositions in English.* Cambridge: Cambridge University Press.

From pre-owned printers to pristine Porsches: A corpus linguistic analysis of item descriptions on eBay

Andrew Kehoe & Matt Gee (Birmingham City University)

This paper presents the first large-scale corpus linguistic analysis of the popular online auction site eBay. Founded in 1995, eBay provides registered users with a marketplace for the sale of a wide range of goods. The site has 152 million active users worldwide, with 800 million items listed for sale at any given time (statistics from eBay Inc. 2014). Although it is often thought of as an auction site where members of the public can sell unwanted gifts and household clutter to the highest bidder, eBay contains an increasing proportion of fixed price (non-auction) items, and 80% of all items are new rather than used. These include goods offered for sale by small businesses using eBay as their main 'shop window', as well as large retailers using the site as additional online sales channel.

Items on eBay are listed in categories, of which there are 35 at the top level ('Computers/ Tablets & Networking', 'Sporting Goods', 'Baby', etc.), each with its own sub-categories. For this study we have built a corpus of item descriptions across all categories on eBay's UK website (http://www.ebay.co.uk), one of 25 country-specific sites. Compiled over four months using our bespoke web crawling tools, this corpus contains over 200,000 item descriptions totalling 50 million words. We included only completed items (items that had closed at the time of our crawl) and, in addition to the textual descriptions, we recorded item

category and sale price. All textual descriptions have been part-of-speech tagged using TreeTagger (Schmid 1994).

In the first part of the paper we describe the corpus compilation process. We present our initial attempts to compile a general eBay lexicon through the examination of word frequencies across item categories. We have found that there are certain core words relating to eBay processes and protocols which appear across all categories: *item*, *buyer*, *seller*, *payment*, *paypal*, *shipping*, *feedback*, etc. Furthermore, we have found that a high concentration of these words tends to be indicative of boilerplate: standard text that appears across multiple listings by different sellers. By identifying boilerplate at an early stage, we are able to focus on more interesting linguistic examples in the remainder of the paper.

Our primary focus, however, is on linguistic differences between the descriptions of items in the various eBay categories, with particular reference to the adjectives used by sellers to describe items for sale in these categories. For this, we adopt a keywords approach (Scott 1997) to compare sub-corpora (categories) with one another. Although the core eBay-related words appear consistently across categories and while there are obvious differences in the frequent topic-related words in each category (primarily nouns), we find significant differences in adjective use between categories too. One example is the words used to describe used items, which vary from *used* itself to *second-hand*, *pre-owned*, *pre-loved*, etc. Another example is items not produced by the usual manufacturer. While the word *fake* is used in some categories, others (e.g. Computers) contain several euphemisms: *non-original*, *generic*, *compatible*, etc. We investigate these in depth, drawing examples from the corpus and carrying out collocational analyses.

A further dimension in our analysis is price. We have produced a price distribution for all items in our corpus and carried out a keyword comparison between the cheapest 25% of items and the most expensive 25%. We present findings from this analysis, including a discussion of the adjectives associated more frequently with items in the cheap (e.g. *lightweight*, *plastic*, *acrylic*, *ex-library*) and expensive (e.g. *heavy*, *steel*, *leather*, *pristine*) categories.

Although there are commercial companies analysing general trends on eBay (e.g. http://terapeak.com), we are not aware of any in-depth academic analyses of the language of eBay or other e-commerce sites. In our paper we give examples of how corpus linguistic techniques can be applied to the study of this increasingly important social phenomenon, and suggest how our techniques could be used to improve the indexing and search functions on sites like eBay.

References

eBay Inc. 2014. *eBay Marketplace Fast Facts At-A-Glance (Q3 2014) – Shareholders' Report:* http://investor.ebayinc.com/common/download/download.cfm?companyid=ebay&fileid=767731&filekey=1b0188fd-9cf4-4dd7-afff-673776dae40d&filename=MP_Factsheet_Q1_2014.pdf

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

Scott, M. 1997. PC analysis of key words – and key key words. *System*, 25(1): 1-13.

WELL as a pragmatic discourse marker of coherence and involvement

John M. Kirk (Technische Universität Dresden)

In the SPICE-Ireland Corpus (Kirk et al. 2011; Kallen & Kirk 2012), which comprises 626,597 words and 15 discourse situations, and which annotates phonological, lexical and syntactical items which function as pragmatic discourse markers (pmw), there are 1763 (pmw 2814) occurrences of WELL. Those occurrences are 22.4% fewer than those in ICE-GB (N=2273/pmw=3565); and 43.9% fewer than those in LLC (N=3143; pmw=6286) (figures from Aijmer 2013). This paper seeks to explain those occurrences and apparent decline in frequency in SPICE-Ireland. To this end, it will review the extensive scholarship on WELL as a pragmatic discourse marker (PDM) from Svartvik (1980) to Aijmer (2013); profile WELL structurally and functionally, provide qualitative as WELL as quantitative results, and draw numerous critical conclusions.

The structural profile deals with pronunciations of WELL as full or rapid; with the tone movement on vocalic nucleus of WELL as a fall, rise, rise-fall, fall-rise, or high plateau; with pauses accompanying WELL before, after, and before & after it; with WELL as an intonation unit on its own, or within an intonation unit shared only with other PDMs, or as part of a syntactic construction; with WELL occurring in utterances in initial, medial or final turn positions, or shared in those positions with other PDMs; and with WELL occurring within declarative, interrogative or imperative constructions.

The functional profile deals with two overarching pragmatic uses of WELL: the marking of discourse coherence, and the marking of speaker involvement. For coherence, pragmatic uses of WELL mark a speaker's search for a particular word or expression or even the content of what it is they want to say; they mark self-repair; or transitions to a new turn or topic, especially in narrative contexts; pragmatic uses of WELL also reflect transitions arising from protocol in institutional settings; they also mark transitions to quotations or citations of direct speech. Finally pragmatic WELL can also functions as a prompt.

For marking speaker involvement with topics, pragmatic uses of WELL mark agreement or disagreement between speakers. In responses to questions or preceding remarks, pragmatic uses of WELL can signal direct engagement with the topic in question or indeed avoidance thereof. They can also mark politeness, negative as WELL as positive, in maintenance and avoidance of threat to face.

The structural and functional profiles combine to describe WELL in SPICE-Ireland with findings for WELL in comparable corpora of spoken varieties of world Englishes, and to draw comparisons which not only highlight characteristics local to Ireland but also to the nature and prevalence of universal tendencies for coherence and involvement in spoken Englishes.

References

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

Kallen, J. L. & J. M. Kirk. 2012. *SPICE-Ireland: A User's Guide.* Belfast: Cló Ollscoil na Banríona.
Kirk, J. M., J. L. Kallen, O. Lowry, R. Rooney & M. Mannion. 2011. *The SPICE-Ireland Corpus: Systems of Pragmatic Annotation for the Spoken Component of ICE-Ireland. Version 1.2.2.* Belfast: Queen's University Belfast and Dublin: Trinity College Dublin.
Svartvik, J. 1980. WELL in Conversation. In S. Greenbaum, G.N. Leech, & J. Svartvik eds. *Studies in English Linguistics for Randolph Quirk.* London: Longman. 167-77.

Routines in lexis and grammar: A 'gravity' approach within the International Corpus of English

Christopher Koch (University of Technology Dresden)

Empirical analyses of linguistic variation in international varieties of English have mostly focused on a small selection of features in few national varieties of English. While these in-depth analyses certainly are relevant for describing the linguistic picture of a single post-colonial variety or a regionally close selection of varieties, this study aims at a more abstract and comparative analysis of a large group of (post-colonial) varieties of English. Building on all components of the *International Corpus of English* (ICE) that have yet been released, it attempts to take on a bird's-eye perspective on unity and diversity at the lexis-grammar interface in World Englishes. This paper focuses on the 'grammar end' of the word-to-word-sequence continuum, studying *n*-grams and colligation as two perspectives on co-occurrence phenomena in lexicogrammar, where "the regular meets the chaotic" (Schneider 2007: 86). The analysis is based on frequency only and abstracts away from individual occurrences in order to guarantee quantifiability and comparability of the results with those on other linguistic levels. This will allow a clustering of the individual varieties (and subgroups thereof, most importantly spoken vs. written data) according to shared and divergent preferences regarding routine phenomena in language use. Finally, it will be demonstrated to which extent the results of this study can be mapped onto established models of varieties of English, most central of which is Schneider's (2003, 2007) dynamic model.

The paper applies two comparatively recent additions to the pool of procedures for the analysis of lexical and, by extension, grammatical co-occurrence patterns: firstly the concept of *lexical gravity* developed by Daudaravičius & Marcinkevičiene (2004) and expanded by Gries & Mukherjee (2010), which offers an appealing alternative to problematic *MI*- or *t*-score based evaluations of collocational attraction (Gries 2010), takes type frequencies into account and leaves *n* variable (cf. also Gries et al. 2011); and secondly *lexical stickiness* introduced by Gries & Mukherjee (2010), which evaluates the preference of any single word to occur in a multi-word sequence. These two measures can be fruitfully combined to estimate variety-specific preferences in routinized language in a border area between lexis and grammar. Using cluster analyses to present the results of the study, this paper shows the possible degrees of "convergence in writing, divergence in speech" (Mair 2007: 84) in varieties of English – setting the 'historical input varieties' of British and American English not as a focal point of the analysis but rather as one point within the larger picture only. This approach will make it possible to depict groups of varieties sharing linguistic preferences,

which will also allow indications towards a *common core* (Quirk et al. 1985) of spoken and written English world-wide.

References

Daudaravičius, V. & Marcinkevičiene, R. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Lingusitics*, 9 (2): 321-48.

Gries, S.Th. 2010. Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. *Proceedings of Corpus Linguistics 2009, University of Liverpool, 20-23 July 2009*. http://ucrel.lancs.ac.uk/publications/cl2009/404_FullPaper.doc.

Gries, S.Th., J. Newman & C. Shaoul. 2011. *N*-grams and the clustering of registers. *Empirical Language Research* 5(1). http://ejournals.org.uk/ELR/article/2011/1.

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 Lingusitics*, 15(4): 520-48.

Mair, C. 2007. British English/American English grammar: convergence in writing-divergence in speech. *Anglia* 125(1): 84-100.

Schneider, E. W. 2003. The dynamics of new Englishes: from identity construction to dialect birth. *Language* 79(2): 233-81.

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

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

"Blessed passion", "pious fever" and "precious death": An analysis of the lexical structure of Early Modern English religious genres

Thomas Kohnen (Cologne University)

Studies on the language of religion have suggested that both "sacred" and "secular" words contribute to the lexical structure of the religious register. But so far there is no sufficiently comprehensive and coherent account of the lexis of religious genres that might determine the exact relationship between the "pious" and "ordinary" elements.

For example, in an early study, Crystal and Davy show that religious language may contain "normal" lexical items, but "the structural relationships between lexical items in religious English differ markedly from those in any other variety: the range of permissible synonyms, antonyms, and so on in religious language is very different from that found elsewhere" (1969: 169). In this view, the special character of religious language is not formed by a large number of particularly "pious" or "devout" words but rather by the idiosyncratic combinations of "ordinary" or "secular" terms (for example, *precious death* or *body and blood*). Görlach (1999), on the other hand, arguing against the background of a text-linguistic analysis, claims that the large variety of religious genres only "share their content, but possibly little else" (1999: 143). Thus, he suggests that the language of religious genres is mainly determined by their special content words, that is, a large number of "sacred" items (such as *blessed passion* and *pious fever*).

In my analysis (based on the Sampler Version of the *Corpus of English Religious Prose* (COERP) and WordSmith Tools) I will look at the distribution of the most frequent content

words in the major Early Modern English religious genres. My analysis will focus mainly on three questions:

- Which "religious" and "secular" items occur in the individual religious genres and what is their proportion?
- What are the specific collocations, relationships and patterns formed by these items?
- Is the frequency and distribution of these items and patterns the same in all religious genres or do they form a cline from the core to the more peripheral genres? (Such a gradience has been shown in religious genres for typical "religious" morpho-syntactic features; see Kohnen et al. 2011).

Thus, this paper seeks to contribute to a more coherent picture of the lexical structure of (Early Modern English) religious language and a more comprehensive perspective on the similarities and differences between the religious register and other, secular genres.

References

COERP: *Corpus of English Religious Prose*. http://anglistik1.phil-fak.uni-koeln.de/coerp.html?&L=1

Crystal, D. & D. Davy. 1969. *Investigating English Style*. Harlow: Longman.

Görlach, M. 1999. *English in Nineteenth-Century England.* Cambridge: Cambridge University Press.

Kohnen, T., T. Rütten & I. Marcoe. 2011. Early Modern English religious prose - a conservative register? In P. Rayson, S. Hoffmann & G. Leech eds. *Methodological and Historical Dimensions of Corpus Linguistics*. VARIENG e-journal, Vol 6.

Scott, M. *WordSmith Tools*. (version 5.0) 2010. PC software by. Lexical Analysis Software Ltd. & Oxford University Press.

Acquisition of phrasal complexity in written intermediate learner language: A case study of NP complexity in German learners of English

Rolf Kreyer & Steffen Schaub (Marburg)

The use of syntactic complexity as a measure to gauge language proficiency dates back at least to 1965 when Hunt introduced his 'Mean Length of T-unit'. Since then measures of syntactic complexity have usually been based on the clause unit (Ortega 2003). It is only recently that phrasal complexity has been introduced as an additional means of assessing (advanced) L2 proficiency (see, for instance, Norris and Ortega 2009). With regard to the interplay of clausal versus phrasal complexity, Biber et al. (2011) put forth an especially interesting hypothesis: while beginning learners rely on clausal complexity, the language of intermediate and advanced learners, especially those entering higher education, increasingly develops towards phrasal complexity. Empirical research testing this hypothesis, however, rests on data from advanced learners only (undergraduate university level or higher) (see, for instance, Parkinson & Musgrave 2014); a tendency which holds true for other areas of L2 research as well. The development of complexity in the language of intermediate learners, i.e. on (advanced) high school level, is a more or less neglected research area.

The present paper is an attempt to contribute to this strand of research by providing an empirical, corpus-based study of the acquisition of phrasal complexity in intermediate learner

language. The data are taken from the Marburg corpus of intermediate learner English (MILE), a longitudinal corpus (currently under compilation) of written exam texts produced by German learners of English from grades 9 through 12. The study is based on a sample of ca. 27,000 words produced by five female and five male pupils. The data are true-longitudinal, as the same group of pupils was accompanied during the entire four-year period. The raw data are scanned for 13 features of noun-phrase complexity (four types of premodification and nine types of postmodification). Furthermore, the data are analysed for accuracy by looking at noun-phrase related error categories, such as determiner errors or word order errors. The following research questions are at the center of attention in this study:

- What types of pre-/postmodification in the noun phrase increase in use from grades 9 to 12?
- In how far can errors within the noun phrase point to problems associated with the acquisition of noun phrase structure?
- To what exent (if at all), can noun phrase complexity be regarded as a reliable indicator of learner proficiency in intermediate learner language?
- In how far does the genre required by the task description have an effect on noun phrase structure?
- To what extent do intermediate learners integrate lexical and structural development, for instance with regard to adjectival premodification?

Overall, the study puts to the test Biber et al.'s (2011) claims concerning the early and intermediate stages of foreign language learning, by focusing on a proficiency band that has not been central to learner corpus research so far.

References

Halliday, M. A. K. 1985. *Spoken and Written Language*. Oxford: Oxford University Press.

Hunt, K. 1965. *Grammatical Structures Written at Three Grade Levels.* Champaigne, IL: National Council of Teachers of English.

Norris, J. M. & L. Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: the case of complexity, *Applied Linguistics,* 30.4: 555-78.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing. *Applied Linguistics,* 24.4: 492-518.

Biber, D., B. Gray & K. Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly,* 45.1: 5-35.

Parkinson, J.& J. Musgrave. 2014. Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, 14.2: 48-59.

Complexity effects in the English comparative alternation: Corpus-based evidence

Gero Kunter (Heinrich-Heine-Universität Düsseldorf)

The alternation between synthetic and analytic comparatives in English has seen considerable interest in recent years (see, for instance, Boyd 2007, D'Arcy 2012, Elzinga 2006, Hilpert 2008, Kytö & Romaine 1997, Mondorf 2003; 2009, Szmrecsanyi 2005). While the comparative of some adjectives is almost always formed either synthetically (e.g. *tough – tougher*, *early – earlier*) or analytically (e.g. *just – more just*, *active – more active*), there is also a considerable number of adjectives for which both comparative forms are available (e.g. *costly – costlier*/*more costly*, *vague – vaguer*/*more vague*). Previous research has shown that the alternation is influenced by a multitude of factors from all domains of linguistic description. For example, the analytic comparative has been found to be more probable if the base adjective ends in /li/, or if the adjective is used in an abstract rather than a concrete sense (e.g. *a more full account* vs. *a fuller hotel*).

Mondorf (2003, 2009) argues that the different factors increasing the probability of analytic comparatives can be subsumed by a single, unifying mechanism called "*more*-support". This mechanism is understood as a compensatory strategy that responds to increases in cognitive complexity either of the adjective or of the context of occurrence: due to inherent processing advantages of the analytic comparative, speakers prefer the analytic form in those environments that are cognitively more demanding.

The present paper tests this hypothesis by combining comparative data from the Corpus of Contemporary English (COCA, Davies 2008-) with reaction times from the English Lexicon Project (ELP, Balota et al. 2007). If the preference of the analytic comparative can indeed be understood as a compensatory response to increases in cognitive complexity, adjectives which are difficult to process (and which therefore have longer reaction times) should occur in COCA with a higher proportion of analytic comparatives than adjectives which are easy to process (and which thus have shorter reaction times).

The proportion of analytic comparatives occurring in COCA is analysed using beta-regression, a family of multivariate models that is particularly suitable for proportion data (Grün et al. 2012). Mean reaction times for each adjective from the ELP are used as the main predictor of interest in the model; other phonological and lexical factors known to affect the English comparative alternation are also included as co-variates. The analysis shows a significant effect of reaction times that is compatible with the notion of *more*-support: adjectives with longer reaction times show a higher proportion of analytic comparatives than adjectives with shorter reaction times. This effect is robust even if the effect of the other co-variates is accounted for. The corpus study thus provides empirical support for the hypothesis that speakers respond to complexity differences if the analytic and the synthetic comparatives are both available.

References

Balota, D. A., M. J. Yap, M. J. Cortese, K. A. Hutchison, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson & R. Treiman. 2007. The English lexicon project. *Behavior Research Methods* 39(3): 445-59.

Boyd, J. 2007. *Comparatively Speaking. A Psycholinguistic Study of Optionality in Grammar*. PhD dissertation, University of California, San Diego.

D'Arcy, A. 2012. On being happier but not more happy. Comparative alternation in speech data. *Working Papers of the Linguistics Circle of the University of Victoria;* 22(1), 72-87.

Davies, M. 2008-. *The Corpus of Contemporary American English (COCA): 450 million words, 1990-present*. http://www.americancorpus.org.

Elzinga, D. 2006. English adjective comparison and analogy. *Lingua*, 116: 757-70.

Grün, B., I. Kosmidis & A. Zeileis. 2012. Extended beta regression in R: shaken, stirred, mixed, and partitioned. *Journal of Statistical Software* 48(11): 1-25.

Hilpert, M. 2008. The English comparative. Language structure and language use. *English Language and Linguistics* 12(3): 395-417.

Kytö, M. & S. Romaine. 1997. Competing forms of adjective comparison in Modern English. What could be more quicker and easier and more effective? In T. Nevalainen & L. Kahlas-Tarkka eds. *To Explain the Present. Studies in the Changing English Language in Honour of Matti Rissanen*, 329-52. Helsinki: Société Néophilologique.

Mondorf, B. 2003. Support for *more*-support. In G. Rohdenburg & B. Mondorf eds. *Determinants of Grammatical Variation in English*. Berlin: Mouton de Gruyter, 251-304.

Mondorf, B. 2009. *More Support for* more-*support*. Amsterdam: Benjamins.

Szmrecsanyi, B. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, 1(1). 113-50.

Lexical cohesion in a contrastive perspective (English - German)

Kerstin Kunz (Universität Heidelberg)
Ekaterina Lapshinova-Koltunski (Universität des Saarlandes, Saarbrücken)
José Manuel Martínez Martínez (Universität des Saarlandes, Saarbrücken)
Katrin Menzel & Erich Steiner (Universität des Saarlandes, Saarbrücken)

This talk contrasts lexical cohesion between English and German, reporting findings from a quantitative lexical analysis[1]. After a discussion of lexical cohesion and its role in establishing coherence against the background of existing (monolingual) studies (Halliday/ Hasan 1976, 1980/1985, Martin 1992, Tanskannen 2006), basic *concepts* will be critically discussed and suitable *methods* for an empirical analysis of lexical cohesion will be suggested. Methodological considerations will be followed by a discussion of *systemic contrasts* between English and German relevant for an investigation of lexical cohesion: lexical resources of the two languages and available mechanisms for lexical cohesion. *A group of assumptions* will then be formulated about *instantial/ textual* differences in lexical cohesion. Differences are assumed between the two languages, but also language internally

[1] The work underlying this paper has been carried out in GECCo (http://www.gecco.uni-saarland.de/GECCo/Home.html), funded by the German Research Foundation (DFG) under GZ STE 840/6-1 and 6-2 und KU 3129/1-2 *Kohäsion im Deutschen und Englischen – ein empirischer Ansatz zum kontrastiven Vergleich.* A large sub-part of the corpus can be queried under https://fedora.clarin-d.uni-saarland.de/cqpweb/. The GECCo-corpus consists of register-comparable originals of *spoken* language (approx. 0.4 million words) and register-comparable originals and their translations of *written* language (approx. 1.0 million words), annotated on various linguistic levels.

between different registers, and within those between written and spoken mode in particular (cf. Hawkins 1986: 28ff, Leisi and Mair 2008: 65ff, Leech et al. 2009: 20, 239, Mair 2006:183, König and Gast 2012: 246ff, Hansen-Schirra et al 2012:76ff, Fischer 2013, Kunz et al. in press, Neumann 2013: 106ff, 166ff, 316ff). One of the assumptions postulates a globally stronger registerial differentiation within German compared to English. More specifically, in English texts, and particularly in certain registers, the breadth of variation in instantiated lexis may be less than in German. Another dependent variable is the relative textual importance of highly frequent words, and of core vs. non-core vocabulary. *The types of data* to be investigated are lexical frequency lists and the role of highly frequent words, lexical density (LD) and type-token-relationships (TTR), part-of-speech (POS) profiles, and classifications of instantiated lexis into Romance/Greek loanwords vs. 'native' Germanic words. It will be shown how selected subcorpora vary in terms of these data, and what these types of variation imply for lexical cohesion. Finally, an initial view will be attempted at the length and density of lexical chains of repetition in our corpus as another measure of lexical cohesion.

Preliminary results indicate greater variation within German than within English in terms of various properties of instantiated vocabulary. They also indicate that in spite of a possibly larger systemic vocabulary of English, the textually instantiated vocabulary may well be smaller than that in German texts, but differently so dependent on register and mode. It also appears as if English registers by and large rely more on core lexis than their German registerial counterparts. Some wider implications for understanding German-English contrasts in (lexical) cohesion will be addressed in a final discussion.

References

Fischer, K. 2013 *Satzstrukturen im Deutschen und Englischen. Typologie und Textrealisierung*. Berlin: Akademie Verlag.

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

Halliday, M. A. K. & R. Hasan. 1980/ 1985. *Text and Context: Aspects of Language in a Social-Semiotic Perspective*. Sophia Linguistica VI. Tokyo: Sophia University.[Republished several times elswhere.]

Hansen-Schirra, S., S. Neumann & E. Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German. Series Text, Translation, Computational Processing*. Berlin / New York: Mouton de Gruyter:

Hawkins. J. A. 1986. *A Comparative Typology of English and German. Unifying the Contrasts*. London. Croom Helm:

König, E. & V. Gast. 2012. *Understanding English-German Contrasts. Grundlagen der Anglistik und Amerikanistik*. Berlin: Erich Schmidt Verlag. [3rd, extended edition].

Kunz, K., S. Degaetano-Ortlieb, E. Lapshinova-Koltunski, K. Menzel, E. Steiner, in press. GECCo - an empirically-based comparison of English-German cohesion. In G. De Sutter, I. Delaere & M.-A. Lefer eds. *New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies*. Berlin: Mouton De Gruyter

Leech, G., M. Hundt, C. Mair & N. Smith. 2009. *Change in Contemporary English. A Grammatical Study*. Cambridge: Cambridge University Press.

Leisi, E. & C. Mair. 2008. *Das heutige Englisch: Wesenszüge und Probleme*. 9. Auflage. Heidelberg: Universitätsverlag Winter.

Martin, J. R. 1992. *English Text. System and Structure*. Amsterdam: Benjamins.

Neumann, S. 2013. *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin: Mouton de Gruyter:
Tanskannen, S.-K. 2006. *Collaborating towards Coherence*. Amsterdam: Benjamins.

Diachronic shifts in agreement patterns of collective nouns in 19th-century American English

Alexander Lakaw (Linnaeus University)

English collective nouns and their agreement patterns, as illustrated in (1)–(3) below, have received a great deal of attention in corpus linguistics. Previous research has found evidence of variability within and across the different varieties of English (e.g. Levin 2001, 2006; Depraetere 2003; Hundt 2006, 2009; Bock et al. 2006; Fernández-Pena 2014).

(1) …and the *police has* not yet been aroused from *its* lethargy.
(COHA; 1822; Magazine)

(2) The *police claim* to know where he is, but *they* will not tell.
(COHA; 1894; Newspaper)

(3) "Here *comes* the *Police*! here *they* come!" shouted the boys,…
(COHA; 1859; Fiction)

This paper fills research gaps identified in previous studies. The first is the lack of diachronic research on agreement patterns. Up to now, only two large-scale studies with a diachronic focus exist (Liedtke 1910, Dekeyser 1975). They, however, were conducted before the era of systematically collected corpora. Secondly, conclusions drawn in several investigations relate the varying agreement within the different varieties of English to the lexical characteristics of the collective nouns themselves (e.g. Depraetere 2003: 124; Bock et al. 2006: 101; Levin 2006: 339). This motivates further research with a focus on intra-linguistic factors (e.g. a semantic perspective). Furthermore, there is a need for further research on extra-linguistic factors that could have influenced agreement with collective nouns. In this study, such an approach is realised by an interdisciplinary socio-historical investigation of collective nouns and the concepts they represent combined with a study on the effects of 19th-century normative grammars.

This paper investigates the agreement patterns and concepts of several collective nouns from four semantic domains, which are 1) PUBLIC ORDER (e.g. *police*, *watch*), 2) MILITARY (e.g. *cavalry*, *army*), 3) FAMILY (e.g. *family*, *couple*) and 4) EMPLOYEES (e.g. *staff*, *crew*), by combining historical corpus linguistics, sociolinguistics and socio-historical perspectives. The vast majority of the investigated material is taken from the *Corpus of Historical American English* (COHA). In order to obtain a reasonable sized amount of material that allows quantitative and qualitative analyses, the investigated time frame of this study was limited to the first 100 years available in COHA, i.e. from 1810 to 1910.

Preliminary results indicate that the variability of some collective nouns can indeed be explained by lexical, socio-linguistic and socio-historical factors. The results show that the agreement pattern of *police* changed from being variable towards a preference of plural

agreement, due to changes made in the organisation of the police patrols prior to the 1870s, which resulted in a shift from singularity towards plurality with regards to the public perception of 'the patrolling police officer'. Contrarily, the agreement pattern of *family* suggests highly variable agreement in the early 1800s but that the proportion of plural verbs decreases significantly in the latter decades of the 19th century. Finally, results indicate that even language-internal constraints advocated by 19th-century prescriptivism affected agreement patterns of collective nouns – a finding that highlights the importance of the role normative grammars play in language change processes of the time.

References

Bock, K., S. Butterfield, A. Cutler, J. C. Cutting, K. M. Eberhard & K. R. Humphreys. 2006. Number agreement in British and American English: disagreeing to agree collectively. *Language,* 82(1), 64-113.

Dekeyser, X. 1975. *Number and Case Relations in 19th-Century British English. A Comparative Study of Grammar and Usage*. Antwerpen/Amsterdam: De Nederlandsche Boekhandel.

Depraetere, I. 2003. On verbal concord with collective nouns in British English. *English Language and Linguistics,* 7(1), 85-127.

Fernández-Pena, Y. 2014. *Verbal Agreement with Collectives Taking of-Dependents: A Corpus-Based Analysis*. Paper presented at ISLE3, University of Zürich, August 2014.

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), 206-32.

Hundt, M. 2009. Concord with collective nouns in Australian and New Zealand English. In P. Peters, P. Collins & A. Smith eds. *Comparative Studies in Australian and New Zealand English*: *Grammar and Beyond* Amsterdam: Benjamins, 207-24.

Levin, M. 2001. A*greement with Collective Nouns in English.* Lund: Lund University Press.

Levin, M. 2006. Collective nouns and language change. *English Language and Linguistic*s, 10(2), 321-43.

Liedtke, E. 1910. *Die numerale Auffassung der Kollektiva im Laufe der englischen Sprachgeschichte*. Königsberg: Karg & Manneck.

Functional and syntactic characteristics of the introductory *it* pattern in academic writing by non-native-speaker and native-speaker students

Tove Larsson (Uppsala University)

The introductory *it* pattern, as in *It is important to distinguish between 'effect' and 'affect'*, is a multifaceted tool used by academic writers not only for information-structural purposes, but also, for example, as a means of persuading the reader of the validity of their claims. This paper aims to investigate the interaction of functional and syntactic characteristics of this pattern in academic writing by non-native-speaker (NNS) and native-speaker (NS) students, as outlined below.

The introductory *it* pattern, which is made up of a non-referential *it*, a predicate and an extraposed clausal subject (Quirk et al., 1985:1391), enables writers to comment on the content of the extraposed clause, for example by using adjectives such as *important* in the predicate. However, while the pattern is used frequently by expert academic writers, it has been found to be problematic for learners (Hewings & Hewings, 2002; Römer, 2009). For

example, with regard to its functional distribution (i.e. its use as a hedge: *it is possible that…* or as an emphatic: *it is essential that…*, etc.), learners have been reported to have a tendency to overuse the pattern to make strong claims compared to NS expert writers (Hewings & Hewings, 2002). No previous study has, however, investigated whether this preference extends to all levels of achievement (i.e. across both higher-graded and lower-graded papers) or whether there is a correlation between its functional and syntactic distribution. Since recent research has shown that the relative frequency of the syntactic types of the introductory *it* pattern (e.g. SVC: *it is interesting to…* or SV: *it seems that…*) differs across levels of achievement, as well as across academic disciplines (Larsson, submitted; see Quirk et al., 1985:1392 for an overview of the syntactic types), the question arises whether the functional preferences might follow the same patterns.

In order to shed light on these research issues, the present contrastive and frequency-based study investigates (i) whether there are any differences with regard to the functional distribution of the pattern across levels of achievement, academic disciplines and NNS vs. NS student writing, and (ii) whether there is a correlation between form and function, i.e. whether the function of the pattern can be predicted based on its syntactic form, or vice versa. Building on previous models for functional classification (e.g. Hewings & Hewings, 2002; Groom, 2005), the model developed for the present study includes a typology of functional categories. Unlike previous studies of this kind, however, the results are also compared to the results of a syntactic classification.

The study uses data from ALEC, a recently compiled corpus of learner writing that, in contrast to most learner corpora, allows for investigation across levels of achievement. The study also makes use of a NS-student reference corpus, which is composed of subsets of BAWE and MICUSP. The tokens were extracted using WordSmith Tools (Scott, 2012) and filtered manually. Preliminary results show that there are disciplinary differences in the functional distribution of the pattern. In addition, there is a difference between lower-graded learners, higher-graded learners, and NS students when it comes to the use of certain functional categories such as introductory *it*-hedges (e.g. *it appears that…*). Furthermore, while there seems to be a correlation between form and function for some categories of the introductory *it* pattern, this does not extend to all categories. The findings of the present study will not only lead to a deeper understanding of the uses of the introductory *it* pattern, but will also help facilitate more targeted teaching for students at different levels of achievement.

References

Advanced Learner English Corpus (ALEC). Corpus compiled at Uppsala University in 2013.

Larsson, T. [submitted]. The introductory *it* pattern: A syntactic analysis of non-native-speaker and native-speaker student writing.

British Academic Written English (BAWE). Corpus compiled at the Universities of Warwick, Reading and Oxford Brookes in 2004–2007. http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/

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.
Michigan Corpus of Upper-level Student Papers (MICUSP). Ann Arbor, MI: The Regents of the University of Michigan. Corpus compiled at the University of Michigan in 2009. http://micusp.elicorpora.info/about-micusp
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
Römer, U. 2009. The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, 7 (1), 140-62.
Scott, M. 2012. WordSmith Tools version 6 (computer program). Liverpool: Lexical Analysis Software.

Anti-vapers: The potential productivity of word-initial/-final affixes and combining forms in spoken English

Jacqueline Laws & Chris Ryder (University of Reading, UK)

The verb to *vape* has been named the word of 2014 by Oxford Dictionaries, but the creation of this verb automatically brings with it a family of related neologisms, such as *vaping*, *vaper* and *anti-vaper*. Novel complex words are coined to fill lexical gaps in the language: by adding prefixes like *anti-* and the agentive suffix *-er* to the verb base *vape*, we create a new word which conveys information that otherwise could only be expressed in a much longer expression: *a person who objects to others who use e-cigarettes*. Complex words such as cheer-*ful* are formed from a base and a suffix; others, such as *un*-kind are formed from a base and a prefix; neoclassical elements, such as *aristo-* and *-cosm* are combining forms which can be attached to the beginning and/or the end of bases, such as *aristo*-crat and micro-*cosm*; some of these can function as both combining forms and affixes.

This paper presents the results of an analysis of derivational morphemes in English using a corpus of complex words containing 17,943 types and 1,008,280 tokens extracted from the spoken component of the British National Corpus (BNC). Firstly, a comprehensive set of 835 affixes, based on Stein (2007), was compiled; this masterlist contained 554 word-initial and 281 word-final derivational morphemes in English. The type and token frequencies of all the associated complex words containing these morphemes, together with their Part of Speech, were recorded. The aim of this investigation was to determine the relative potential productivity of the 490 derivational morphemes that occurred in the spoken corpus, of which 48% were pre/suffixes, 21% were word-initial/-final combining forms, and 31% were derivational morphemes that function as both combining forms and pre/suffixes.

Of the master set of 554 word-initial morphemes, 48% occurred in the BNC spoken corpus and, of the 281 word-final morpheme set, 79% were recorded: this robust word-final preference effect was observed with all three affix types (affixes, combining forms and those morphemes that function as both), in line with the well-established phenomenon that there is a cross-linguistic preference for suffixation over prefixation (Greenberg, 1966). More precisely, statistical comparisons revealed that the proportion of the suffix set that appeared

in the corpus was significantly greater than that predicted from the master set of derivational morphemes analysed, and the representation of word-initial combining forms was significantly less than that predicted. The other differences were in keeping with the relative sizes of the two morpheme sets (the masterlist and the observed list).

Productivity, in terms of the number of complex word types that were recorded for each of the three categories of derivational morphemes, was found to be greatest for the prefix-suffix set and furthermore, suffixes were 11% more productive than prefixes; this difference reached statistical significance. This finding accords with the prediction that suffixes would be expected to be more productive than prefixes in English; however, the opposite trend was observed for derivational morphemes that function as both affixes and combining forms: here a significant 12% advantage for the word-initial category was obtained. This paper has three aims: to report the relative distributions of affix types in spoken English, to present the relative productivity of word-initial and word-final affixes as a function of the metrics developed by Baayen and colleagues (Baayen, 2009), and to discuss the implications of these findings in relation to affix position preferences in English.

References

Baayen, R. H. 2009. Corpus linguistics in morphology: morphological productivity. In A. Luedeling & M. Kytö eds. *Corpus Linguistics. An International Handbook.* Berlin: Mouton De Gruyter, 900-19.

Greenberg, J. H. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg ed. *Universals of Language*, 2nd ed. Cambridge, Mass: MIT Press.

Stein, G. 2007. *A Dictionary of English Affixes: Their Function and Meaning*. Munich: Lincom Europa.

Noun-noun compounds and their unpacked variants in the 20th century

Hans Martin Lehmann & Gerold Schneider (University of Zurich)

Over the years compound nouns have received attention from a wide variety of perspectives. It is widely recognized that the packing and unpacking of compounds is a complex process that is not directly predictable by the sum of the parts involved.

As Jespersen (1942:137) notes: “Compounds express a relation between two objects or notions, but they say nothing of the way in which the relation is to be understood. That must be inferred from the context or otherwise.” This is often illustrated by the obvious difference in the unpacking of compounds like *olive oil, baby oil* or *malaria mosquito*. The interpretation *oil made of olives, oil for babies* and *mosquito that transmits malaria* cannot be derived directly from the de-contextualised compound.

In our paper we present a data-driven approach that leverages the new empirical possibilities offered by large sets of data and automatic annotation. We make use of a set of annotated data comprising over 2.5 billion words, from which we extract syntactic patterns in which

participants of compound nouns occur. For each compound we extract verbal and prepositional paraphrases. The compound *malaria mosquito* in (1) is automatically mapped to by our approach to its unpacked form in (2), which renders the relation between *malaria* and *mosquito* explicit. Table 1 shows a list of verbs occurring in the subject-verb-object dependency structure defined by the compound in *malaria mosquito*.

(1) The *malaria mosquito* actually only bites you below your knee. (CNN:20070807lkl.01)

(2) [...] flooding has left vast pools of standing water where *mosquitoes* can *breed* and *spread malaria* and dengue fever. (NYT0613:WORLD)

Table 1. Frequency list of verbs in the mosquito:subject - x:verb - malaria:object frame in 2.5 bn words. n=40

Rank	**n**	**verb**
1	14	*spread*
2	10	*carry*
3	10	*transmit*
4	3	*do*
5	2	*cause*
6	1	*have*

The functional relations include location, purpose, cause, composition, patient, and many more. In order to extract such relations in a text-driven fashion and follow the formation processes, we treat noun-noun sequences as alternations in the sense of Levin (1993). Noun compounds are alternates to PPs (examples (3) and (4)) and to verbal relations (e.g. subject-verb-object as in sentence (2).

Example (4) shows the compound *interest rate*, and (3) its prepositional paraphrase.

(3) The *rate of interest* at which this assistance is provided indicates the Bank's view of the appropriate level of interest rates. (BNC:H8E:95)

(4) [...], firms compare the rate of return on a new investment with the *interest rate* at which they must borrow to finance the project. (BNC:HGP:2082)

We establish the ratio in which lexical types occur in compound and paraphrased forms and discuss the restrictions and factors in the choice involved. We evaluate in how far our approach allows us to model the alternation and automatically disambiguate between the possible underspecified functional relations. This provides us with an envelope of variation (Labov 1969, Sankoff 1988). Machine-learning approaches like Nakov and Hearst (2008) have used related methods.

As Biber (2003: 175-6) has shown, noun-noun sequences have steadily increased over the last four centuries. Leech et al. (2009) have been able to show an increase for a relatively short time-span, between the 1960's and the 1990's.

Beyond providing frequencies for compound nouns, we trace the genesis of now popular noun-noun types in the ARCHER corpus, and we project the established compound - paraphrase relationships into the set of the BROWN family of corpora covering the 20th century to the present. We investigate the proportion of paraphrased variants versus their corresponding compounds in the different text types and the development of that ratio over the observed century of American and British data.

Our approach to modelling noun-noun sequences as alternations may even give us the tools to trace ad-hoc formations and neologisms, based on when unpacked variants start to shift towards packed noun compounds.

References

Biber, D. 2003. Compressed noun-phrase structures in newspaper discourse: the competing demands of popularization vs economy. In J. Aitchison & D. Lewis eds. *New Media Language*. London: Routledge, 169-181.

Jespersen, O. 1942. *A Modern English Grammar on Historical Principles*, Vol. 6. Copenhagen: Munksgaard.

Labov, W. 1969. Contraction, deletion, and inherent variability of the English copula. *Language*, 45(4): 715-62.

Leech, G., M. Hundt, C. Mair & N. Smith. 2009. *Change in Contemporary English. A Grammatical Study*. Cambridge:Cambridge University Press.

Levin, B. C. 1993. *English Verb Classes and Alternations: a Preliminary Investigation*. Chicago: University of Chicago Press.

Nakov, P. & M. Hearst. 2008. Solving Relational Similarity Problems Using the Web as a Corpus. In *Proceedings of ACL-08*, Columbus, Ohio, 452-60.

Sankoff, D. 1988. Sociolinguistics and syntactic variation. In F. J. Newmeyer ed. *Linguistics: the Cambridge Survey*. Cambridge: Cambridge University Press, 140-61.

Lexis in EFL theses

Signe-Anita Lindgrén (Åbo Akademi University)

Two features previously seen to distinguish academic texts from other texts are (i) the use of so called academic vocabulary (AV), such as those collected in academic vocabulary lists (AWL Coxhead, 2002; NAWL Browne, Culligan, & Phillips, 2013) and (ii) readability, as measured, for instance, by the proportion between number of words, sentences, and syllables (e.g. Hartley, 2008). Recent studies have, however, questioned the usefulness and relevance of academic word lists pointing, for instance, to discipline specific variation, and of quantitative-structural measures (e.g. Hyland & Tse, 2007, Hartley, 2008). This study (i) explored whether these features can - despite this - still be successfully employed to distinguish between specific texts written by advanced EFL writers, (ii) asked whether the use might be reflected in the evaluation of the texts, and (iii) discussed to what extent it is adequate to draw the writers' attention to these features.

The data for the study consisted of BA and MA theses written in English by EFL students of English language and literature with the same L1 (Finland-Swedish; the BATMAT Corpus).

The corpus contained approx. 2.5 million words and comprised theses written on literary vs. linguistic topics. The academic vocabulary lists of Coxhead (2002) and Browne et al. (2013), as well as the Flesh reading ease test (cf. Flesch 1951; e.g. Hartley, 2008) were employed. Prior to the quantitative analyses, the parts of the texts that were considered not main body of the text and language originally produced by the writers were removed: tables of contents, bibliographies, appendices, direct quotes within the text, summaries in the native language and such. The following non-parametric tests were applied: the Kruskal-Wallis test, the Mann-Whitney test, the Wilcoxon signed-rank test, Spearman's r, and Rosenthal's effect size measure.

The study thus investigated the frequency use of academic words and to what degree these texts exhibited readability scores that are considered "academic" (Flesch 2007; cf. Flesch 1951). The study further contrasted theses of different proficiency levels (BA vs. MA), and of different types, i.e. thesis written on linguistically-oriented vs. literary-oriented topics, with respect to these features. Moreover, it answered the question as to how the use of these two features correlated statistically with grades awarded by external raters. As both the AWL and the NAWL were used, the study also allowed for comparisons of results between these two word lists. The results showed interesting results both with regard to the two levels and with regard to theses written on linguistic topics in contrast to literary topics.

The paper briefly presents previous studies, and the current data and research methods, and then focuses on the findings. In addition to a general interest in describing specific features of advanced non-native writing, exploring the use, or the relevance of academic word lists and the readability score, and comparing the two vocabulary lists, the findings also have direct pedagogical implications.

References

Coxhead, A. A new academic word list. *TESOL Quarterly* 34 (2002): 213-38.
Browne, C., B. Culligan & J. Phillips. 2013. The new academic word list. http://www.newacademicwordlist.org/.
Flesh – Document Readability Calculator. Flesh. 2007. http://flesh.sourceforge.net/.
Flesch, R. 1951. *How to Test Readability*. New York: Harper & Brothers.
Hartley, J. 2008. *Academic Writing and Publishing.* Taylor & Francis e-Library.

Critical issues in spoken corpus development: Defining a transcription scheme for the spoken BNC2014

Robbie Love (Centre for Corpus Approaches to Social Science (CASS), Lancaster University)

The Centre for Corpus Approaches to Social Science (CASS) at Lancaster University and Cambridge University Press (CUP) are collaborating on a new corpus of spoken British English, known as the Spoken British National Corpus 2014 (Spoken BNC2014). This will be the first publicly-accessible corpus of its kind since the spoken component of the original British National Corpus (Leech 1993) (henceforth Spoken BNC1994).

As with all spoken corpora, the definition of a suitable transcription scheme protocols not only an essential preparatory step, but also a locus for the critical examination of certain issues in corpus construction. The Spoken BNC2014's transcription scheme was developed cooperatively by CASS and CUP via a process of reviewing previous work, pilot-testing, and extensive discussion/consultation. We could in theory have reused, unedited, the same scheme as Spoken BNC1994 (Crowdy 1994); however, we argue that this scheme is insufficiently detailed to minimize ambiguity in transcription, as too much is left to the discretion of individual transcribers. Instead, we adapted a modern, highly-detailed system, one currently being used in another CASS project (Gablasova et al., under review), while also, in the interests of comparability, taking into account Crowdy (1994) where possible.

We refined the transcription scheme for the Spoken BNC2014 via the following steps:

(1) Beginning with the scheme of Gablasova et al. (under review), designed for formal, one-to-one learner data, we removed all features deemed irrelevant to spontaneous conversations with varying numbers of speakers.
(2) We critically evaluated the specific recommendations of Atkins et al. (1992) and the practices outlined in Crowdy (1994), modifying the scheme accordingly; key issues here were *encoding speaker IDs*, the role of *anonymization* (drawing upon Hasund 1998), the marking of *overlaps*, the optimal approach to *punctuation*, and the previously unconsidered possibility of encoding *quotative speech*.
(3) We produced a working draft of the thus-modified scheme, which we then piloted with two full-time professional transcribers at CASS, using the first tranche of audio recordings collected for the corpus.
(4) We evaluated the pilot via ongoing face-to-face consultation with the transcribers, and at the end by analysis of the output. The pilot test confirmed that it is feasible to anonymize speakers during transcription, as well as giving transcribers freedom to use intuitive criteria for the coding of some at-issue features such as question marks. However, it is unlikely to be feasible to mark quotative content consistently enough to be worthwhile. Moreover, we identified problematic methodological issues with regards to *speaker identification*.
(5) Prior to implementation, the scheme was amended in light of the pilot and of further analysis of at-issue features including *non-standard words/sounds, non-English speech, pauses,* and *events*.

The resulting scheme adheres to no particular prior encoding standard. However, its coding is defined so as to be computationally unambiguous, allowing automated mapping to standard XML for distribution/archiving; our target encoding instantiates "Modest XML for Corpora" (Hardie 2014). A critical comparison of this scheme to that of the Spoken BNC1994 explores the delicate balance we sought between backwards-compatibility and optimal practice in the context of the new corpus.

References

Atkins, A., Clear, J., & Ostler, N. 1992. Corpus design criteria. *Literary and Linguistic Computing, 7*1: 1-16.

Crowdy, S. 1994. Spoken corpus transcription. *Literary and Linguistic Computing, 9*1: 25-28.
Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. [under review] Epistemic stance in spoken L2 English: The effect of task type and speaker style. Submitted *to Applied Linguistics*.
Hardie, A. 2014. Modest XML for corpora: not a standard, but a suggestion. *ICAME Journal, 38*: 73-103.
Hasund, K. 1998. Protecting the innocent: the issue of informants' anonymity in the COLT corpus. In A. Renouf, ed. *Explorations in Corpus Linguistics.* Amsterdam: Rodopi, 13-28.
Leech, G. 1993. 100 million words of English. *English Today*, 9, 1: 9-15. doi:10.1017/ S0266078400006854.

Tracing the variation between *perhaps* and *maybe* in historical and contemporary corpora

María José López-Couso (University of Santiago de Compostela)

The notions of evidentiality and epistemicity have recently attracted considerable attention among scholars working in different frameworks (see, for example, the 2009 special issue of *Functions of Language* and Marín-Arrese et al. 2013). In this context, over the last few years I have been involved in a joint research project focusing on the origin, development, and present-day use of epistemic/evidential expressions in English, which has so far dealt with the parentheticals *it seems* (López-Couso & Méndez-Naya 2014a) and *looks like* (López-Couso & Méndez-Naya 2014b) and with the emergence of the epistemic adverb *maybe* (López-Couso & Méndez-Naya 2014c, submitted). This paper represents a further step in this project by focusing on the variation between *perhaps* and *maybe*, two near synonymous epistemic adverbs conveying weak modality (Huddleston & Pullum et al. 2002: 769), more specifically, some degree of doubt towards the truth of the speaker's proposition (Quirk et al. 1985: 620).

The first part of my presentation discusses the processes of grammaticalization which gave rise to the adverbs *perhaps* and *maybe*, thus providing the background for the variationist analysis in the second part of the paper. Evidence for this historical account is drawn from the standard historical dictionaries (OED, MED), the Helsinki Corpus, and ARCHER 3.2. *Perhaps* originates in the prepositional phrase *per haps*, a combination of the Latin or Anglo-Norman preposition *per* 'for, by' and the Scandinavian noun *hap* 'occurrence, chance' (see OED s.v. *perhaps*, adv. and n.). In turn, *maybe* goes back to a complementation structure of the type *it may be* (*that*)..., which evolved into a parenthetical, and finally into an adverb (see López-Couso & Méndez-Naya 2014a; submitted).

Given that the historical evidence shows that the 20th century is the key period in the competition between *maybe* and *perhaps*, the second part of my presentation focuses on the variation between the two adverbs over the last half century. For the purposes of this paper, I rely on data from six corpora belonging to the Brown family: LOB, FLOB and BE06 for British English, and Brown, Frown and AmE06 for American English. The detailed analysis of the material provides robust quantitative evidence for the similarities and differences between the two adverbs as regards (i) their syntactic behaviour (e.g. the position they occupy in the clause and the syntactic function they realize); (ii) their degree of formality and

their distribution across text-types (e.g. is *maybe* more informal than *perhaps*, as suggested in Quirk et al. (1985: 620) and in dictionaries of contemporary English?); (iii) their distribution in the two major varieties of English (e.g. is *maybe* more common in American English than in British English, as mentioned in Biber et al. (1999: 868)?); and (iv) potential diachronic changes in the use of the two adverbs over the last fifty years (e.g. is there any evidence of recent or ongoing processes of colloquialization and Americanization; see Leech et al. (2009)?).

References

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

Ekberg, L. & C. Paradis eds. 2009. Special Issue on Evidentiality in Language and Cognition. *Functions of Language*, 16/1.

Huddleston, R.; G. Pullum et al. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Leech, G., M. Hundt, C. Mair & N. Smith eds. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

López-Couso, M. J. & B. Méndez-Naya. 2014a. 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. & B. Méndez-Naya. 2014b. 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. & B. Méndez-Naya. 2014c. On the adverbialization of *may* + *be/happen* constructions. Paper presented at ICAME 35, Nottingham 30 April - 4 May 2014.

López-Couso, M. J. & B. Méndez-Naya. [Submitted.] From clause to adverb: On the history of *maybe*.

Marín-Arrese, J., M. Carretero, J. Arús Hita & J. van der Awera eds. 2013. *English Modality: Core, Periphery and Evidentiality*. Berlin: De Gruyter Mouton.

MED = Kurath, H. et al. 1952-2001. *Middle English Dictionary*. Ann Arbor: University of Michigan Press. http://ets.umdl.mich.edu/m/med/.

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

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

Co-occurrence and iteration of intensifiers in early English: A case of linguistic accretion?

Belén Méndez-Naya (University of Santiago de Compostela)

Intensifying adverbs indicate the degree or the exact value of the item they modify (e.g. *very happy*, *absolutely adore*). Like other degree expressions, such as comparatives and superlatives, in Present-day English intensifiers co-occur only very exceptionally (Huddleston & Pullum et al. 2002: 585), if at all (Paradis 2001: 55). Thus, for instance, the *Corpus of Contemporary American English* (450 million words) does not contain any examples of the string **extremely very* + adjective* and just six of the sequence **very extremely* + adjective* (cf. (1)). Somewhat more frequent are cases in which degree adverbs

iterate, as in (2) (see, e.g. Bolinger 1972: 290; Paradis 1997: 10; Huddleston & Pullum et al. 2002: 585).

(1) then I didn't get the money and I got a **very extremely** violent beating for that. (COCA, 2008, SPOK)

(2) and When she was good she was **very very** good and when she was bad she was horrid. (COCA, 2007, FIC)

Co-occurrence and iteration of intensifiers does not seem to have been so constrained in earlier English (Méndez-Naya 2007), as evinced by instances such as (3) and (4).

(3) & þurrh þa seffne innseȝȝless wass Rihht swiþe wel bitacnedd þatt sefennfald godleȝȝc þatt Crist Uss dide þurrh hiss come; (HC, QM1_IR_HOM_ORM)

(4) ða ic ða ðis eall gemunde, ða wundrade ic swiðe swiðe ðara godena wiotona ðe giu wæron giond Angelcynn (HC, QO2_XX_PREF_PRCP)

It could be argued that intensifying structures of the type shown in (1)-(4) represent cases of a common cross-linguistic phenomenon variously referred to in the literature as "accretion" (Kuteva 2008), "hypercharacterization" (Lehmann 2005), and "pleonasm" (Lehmann 2005; Sornicola 2006). Accretion involves the accumulation of "redundant" linguistic material and has been shown to play an important role in the development of new grammatical structures (López-Couso 2013, 2014).

In this paper I study co-occurrence and iteration patterns with five intensifiers in the early stages of the English language, namely *swiþe*, *full*, *right*, *well*, and *very*. The rationale behind this selection lies, above all, in their high frequency of use: they are the five most common intensifiers in the Old, Middle, and Early Modern English periods. Moreover, the selected intensifying adverbs have been shown to illustrate longitudinal change, since "old", "worn-out" intensifiers tend to be replaced by other more expressive items over time (cf. Bolinger 1972: 18).

My presentation pays attention to the following aspects: (i) the similarities and differences between intensifier repetition and co-occurrence and cases of simple modification; (ii) the function(s) of these combinations (e.g. are they used for emphasis or to compensate for the loss of expressivity of a high-frequency intensifier?); (iii) the potential factor(s) explaining the order of accrued intensifiers (e.g. age and semantic load of the intensifiers involved; association of the left-most position with more grammaticalized values (cf. Adamson 2000); and (iv) adjective/verb type (i.e. does adjective/verb type play a role in the accretion of intensifiers?). Given that intensifier co-occurrence and intensifier iteration are low-frequency phenomena, for the purposes of this paper the Penn corpora (YCOE, PPME2, and PPCEME) were used as a base line, supplemented (whenever necessary) with additional data from complementary sources, including the OED, the MED, the *Corpus of Early English Medical Writing*, and the *Dictionary of Old English Corpus*.

References

Adamson, S. 2000. A lovely little example: word order options and category shift in the premodifying string. In O. Fischer, A. Rosenbach & D. Stein eds. *Pathways of Change. Grammaticalization in English*. Amsterdam: Benjamins, 39-66.

Bolinger, D. 1972. *Degree Words*. The Hague & Paris: Mouton.

Huddleston, R. & G. Pullum et al. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Kuteva, T. 2008. On the frills of grammaticalization. In M. J. López-Couso & E. Seoane eds. *Rethinking Grammaticalization. New Perspectives*. Amsterdam: John Benjamins, 189-217.

Lehmann, C. 2005. Pleonasm and hypercharacterisation. In G. Booij & J. van Marle eds. *Yearbook of Morphology 2005*. Dordrecht: Springer, 119-154.

López-Couso, M. J. 2013. Exploring linguistic accretion: Middle English as a testing ground. Plenary lecture, *8th International Conference on Middle English (ICOME-8)*, Murcia, 2-4 May 2013.

López-Couso, M. J. 2014. On structural hypercharacterization: Some examples from the history of English syntax. Plenary *18th International Conference on English Historical Linguistics*, Leuven, 14-18 July 2014.

Méndez-Naya, B. 2007. *He nas nat right fat*: On the origin and development of the intensifier *right*. In G. Mazzon ed. *Studies in Middle English Forms and Meanings*. Frankfurt am Main: Peter Lang, 191-207.

Mustanoja, T. F. 1960. *A Middle English Syntax*. Helsinki: Société Néophilologique.

Paradis, C. 1997. *Degree Modifiers of Adjectives in Spoken British English*. Lund: Lund University Press.

Paradis, C. 2001. Adjectives and boundedness. *Cognitive Linguistics*, 12(1): 47-65.

Sornicola, R. 2006. Expletives and dummies. In K. Brown ed. *Encyclopedia of Language and Linguistics, 2nd ed., vol.* 4. Oxford: Elsevier, 399-410.

On the relation of corpus data and eWAVE: Pluralized mass nouns in East and West African Englishes

Susanne Mohr (University of Bonn)

For the typological study of the varieties of English, the *(electronic) World Atlas of Varieties of English* (eWAVE) (Kortmann & Lunkenheimer 2013) is one of the most important tools of research. The electronic database comprises 235 morphosyntactic features including frequency ratings of these in each variety. However, the ratings are not always based on empirical results.

In the past decades, especially the "New" Englishes spoken in Asia and Africa, have received increased attention (e.g. Platt et al. 1984; Kachru et al. 2006; Mesthrie 2008). Typologically, they have shown similar traits (e.g. Szmrecsanyi & Kortmann 2009). For Africa, Huber (2012) mentions four morphosyntactic features that are typical: F55 – different count/mass distinctions resulting in use of plural for StE singular, F59 – double determiners, F166 – *come*-based future/ingressive markers and F222 – *too; too much; very much* "very" as qualifier.

The paper investigates the first of the four features, the non-standard use of plural mass nouns, and its frequency in two East (Kenyan and Tanzanian) and two West (Nigerian and Ghanaian) African varieties of English. It takes the frequency ratings of eWAVE as a starting

point and evaluates these with respect to corpus data from the African components of the *International Corpus of English (ICE)* and the African components of the *Corpus of Global Web-Based English (GloWbE)* (Davies 2013). The *British National Corpus* was used as a reference corpus. 21 mass nouns belonging to three different categories were investigated:

a. Nouns that cannot be inflected for plural under any circumstances, such as *furniture*
b. Nouns that can be pluralized in special circumstances, such as *water* in *to navigate the waters of the Indian Ocean*
c. Flexible terms such as *tea* in *she likes tea* and *the teas of India are very tasty*

Mohr (2014) showed that the claimed high frequency of pluralized mass nouns in Kenyan English in eWAVE seems not applicable given a corpus analysis of *ICE – East Africa* (Schmied et al. 1999) and *GloWbE*. Data from the West African corpora further show that in West African Englishes, the frequency of plural mass nouns is higher: *luggage*, for instance, is not pluralized at all in the East African data, while approximately one third of all instances of the term in ICE-Nigeria showed plural markings. Similarly, *offsprings* is always pluralized in ICE-Ghana[1] but only in one third of the cases in ICE-EA. Chi-square tests also revealed that many of the differences between West African and British English are statistically significant, while the East-African British differences are not. Several of the mass nouns show semantic change compared with Standard British English. An example is *silverwares*, of which all tokens in the Kenyan data refer to (sports) trophies.

The abovementioned categories of nouns proved to be important in terms of possible explanations of the phenomenon. A distinction of mass nouns into subject (b above) and object-mass nouns (a above) that show differences in the mapping of semantics and grammar in that subject-mass nouns refer to non-countable entities while object-mass terms refer to countable entities and both cannot be inflected for plural, seems to be decisive (Wiese 2012). This is in line with the results from Mohr (2014).

Ultimately, the paper is an appeal to re-evaluate eWAVE as an important typological tool in World Englishes research, and to verify the frequencies of the listed features by empirical results. Further, it questions the typological grouping of African Englishes based on region alone.

References

Davies, M. 2013. *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. http://corpus2.byu.edu/glowbe/. Accessed 07-09-2014.

Huber, M. 2012. Regional profile: Africa. In B. Kortmann & K. Lunkenheimer eds. *The Mouton World Atlas of Variation in English.* Berlin/Boston: Mouton de Gruyter, 806-23.

Kachru, B. B., Y. Kachru & C. L. Nelson eds. 2006. *The Handbook of World Englishes*. Oxford: Blackwell.

Kortmann, B. & K. Lunkenheimer eds. 2013. *The Electronic World Atlas of Varieties of English*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://ewave-atlas.org. Accessed 28-07-2014.

[1] A preliminary version of the corpus was kindly provided by Magnus Huber at the University of Gießen. The results of the analysis are preliminary as well.

Mesthrie, R. 2008. *Varieties of English: Africa, South and Southeast Asia.* Berlin: Mouton de Gruyter.
Mohr, S. 2014. *Informations* on *researches* and *knowledges* – mass nouns in East African Englishes. Paper presented at the *3rd Bonn Applied English Linguistics Conference* (BAELc3), Bonn.
Platt, J. T., H. Weber & Ho Mian Lian. 1984. *The New Englishes*. London: Routledge & Kegan Paul.
Schmied, J., D. Hudson-Ettle & B. Krohne. 1999. *The International Corpus of English – East Africa*. Bayreuth/Chemnitz.
Szmrecsanyi, B. & B. Kortmann. 2009. Vernacular universals and angloversals in a typological perspective. In M. Filppula, J. Klemola & H. Paulasto eds. *Vernacular Universals and Language Contacts.* New York/London: Routledge, 33-53.

Phraseological patterns in corporate environmental discourse: A corpus-driven diachronic study

Alessandra Molino (University of Turin, Italy)

This paper deals with corporate discourse on climate change and environmental sustainability. The interest in the way companies articulate their commitment to environmental protection has grown steadily in recent years in various fields, such as economic management, policy studies and communication science (Schlichting 2013). Corporate sustainability has also attracted the attention of discourse analysts (Coupland 2005; Bowers 2010; Catenaccio 2011; Caimotto and Molino 2011; Fuoli 2012; Caimotto 2013; Lischinsky and Egan Sjölander 2014); however, there have been relatively fewer corpus studies of environmental disclosures (see Lischinsky 2010, 2011, 2014; Malavasi, 2011, 2012; Molino, 2013). Nevertheless, corpus linguistics approaches seem particularly appropriate to investigate the resources employed by companies to frame discourses of environmental sustainability, a topic which deserves investigation given the ability of businesses to "influence public dialog and understanding of ecological issues" (Pal and Jenkins 2014: 389).

The aim of this paper is to provide empirical evidence of the way businesses construct meaning in their environmental disclosures. In particular, this study seeks to map whether and how corporate environmental discourse has changed since the early 2000s, following existing studies (Bowers 2010; Schlichting 2013) suggesting that shifts have occurred from a rather sceptical view of the risks of climate change in the early 2000s to the acknowledgment that global warming is happening and the idea that companies should become leaders in finding solutions to environmental problems (Pal and Jenkins 2014).

Assuming that meaning is constructed not only through words but, most crucially, through the phraseological patterns in which words occur, this study analyses recurrent co-occurrence structures, both lexical and grammatical, involving as node words the most frequent keywords extracted from each year of a diachronic corpus of corporate disclosure documents published from 2000 to 2014. The corpus consists of texts taken from official reports by ten of the highest ranking "clean capitalism leaders", as indicated in Corporate Knights' "Global 100 Most Sustainable Companies in the World" (2006-2014) (Corporate Knights 2014). A variety of publications were consulted, such as Sustainability Reports and Integrated Annual

Reports, and only the sections specifically dealing with the environment were saved as plain text files and included in the corpus.

The research questions addressed in this study are: 1) Have keywords related to environmental sustainability in the discourse of large corporations changed since year 2000? 2) Have the phraseological patterns associated with the most recurrent keywords changed over time and how? 3) Do corpus data suggest the existence of boundary points in time when rhetorical shifts have occurred in the way businesses discursively construct the idea of environmental sustainability? 4) Do corpus results corroborate the picture provided in qualitative, non-language focused diachronic studies of corporate environmental discourse (e.g. Bowers 2010; Schlichting 2013)?

References

Bowers, T. 2010. From image to economic value: a genre analysis of sustainability reporting, *Corporate Communications: An International Journal* 15 (3): 249-62.

Catenaccio, P. 2011. Social and environmental reports: a diachronic perspective on an emerging genre. In G. Garzone & M. Gotti eds. *Discourse, Communication and the Enterprise. Genres and Trends*. Bern: Peter Lang, 169-92.

Caimotto, M. C. 2013. The unsustainable Anglicization of sustainability discourse in Italian green companies. In R. Salvi & W. Cheng eds. *Textus* XXVI (1). Carocci: Roma. 115-26.

Caimotto, M. C. & A. Molino 2011. Anglicisms in Italian as alerts to greenwashing: a case study. In *Critical Approaches to Discourse Analysis Across Disciplines (CADAAD).* 5 (1). 1-16.

Coupland, C. 2005. Corporate social responsibility as argument on the Web. *Journal of Business Ethics*, 64 (2): 355-66.

Corporate Knights (2014. Global 100 Media Kit, http://global100.org/wordpress/wp-content/uploads/2014/01/G100_Media-Kit.pdf. Accessed 14 Dec 2014.

Fuoli, M. 2012. Assessing social responsibility: a quantitative analysis of appraisal in BP's and IKEA's social reports. *Discourse & Communication* 6(1), 55-81.

Lischinsky A. 2010. The struggle over sustainability: a corpus approach to managerial conceptions of sustainable development. In *Proceedings of the Critical Approaches to Discourse Analysis Across the Disciplines Conference*, 13-15 Sept, Łódź, Poland.

Lischinsky, A. 2011. The discursive construction of a responsible corporate self, in tracking discourses: politics, identity and social change. In A. Egan Sjölander, J. G. Payne eds. *The Discursive Construction of a Responsible Corporate Self*. Nordic Academic Press Editors. 257-85.

Lischinsky, A. 2014. What is the environment doing in my report? Analysing the environment-as-stakeholder thesis through corpus linguistics. *Environmental Communication. A Journal of Nature and Culture*, doi: 10.1080/17524032.2014.967705.

Lischinsky A. & A. Egan Sjölander. 2014. Talking green in the public sphere: press releases, corporate voices and the environment. *Nordicom Review* 35 [Special Issue]. 125-39.

Malavasi, D. 2011. "Doing well by doing good": a comparative analysis of Nokia's and Ericsson's corporate social responsibility reports. In G. Garzone & M. Gotti eds. *Discourse, Communication and the Enterprise. Genres and Trends*. Bern: Peter Lang, 193-212.

Malavasi, D. 2012. 'The necessary balance between sustainability and economic success': an analysis of Fiat's and Toyota's corporate social responsibility reports. In P. Heynderickx, S. Dieltjens, G. Jacobs, P. Gillaerts & E. de Groot eds. *The Language Factor in International Business: New Perspectives on Research, Teaching and Practice*. Bern: Peter Lang, 247-64.

Molino A. 2013. 'New targets' for 'more sustainable' companies: a corpus-driven study of Adidas', Ikea's and Vodafone's sustainability reports. In R. Salvi & W. Cheng eds. *Textus* XXVI (1). Carocci: Roma. 103-13.

Pal, M. & Jenkins, J. 2014. Reimagining sustainability: an interrogation of the Corporate Knights' Global 100, *Environmental Communication* 8(3): 388-405.
Schlichting, I. 2013. Strategic framing of climate change by industry actors: a meta-analysis, *Environmental Communication* 7(4): 493-511.

A spoken corpus of Cameroon Pidgin English

Gabriel Ozon (Sheffield)
Melanie Green (Sussex)
Miriam Ayafor (Yaounde)

Cameroon Pidgin English (CPE) is an expanded pidgin/creole spoken in some form by an estimated 50% of Cameroon's 22,000,000 population (Lewis et al. 2014), primarily in the Anglophone west regions, but also in urban centres throughout the country. As a primarily spoken language, CPE has no standardised orthography, but enjoys a vigorous oral tradition, not least through its presence in the broadcast media. However, it resists close documentation due to its stigmatised status in the face of French and English, prestige languages of Cameroon, where it also co-exists with an estimated 280 indigenous languages (Lewis et al. 2014). The majority of publications on English in Cameroon have to date focused mainly on the sociolinguistic aspects, with little close grammatical detail (e.g., De Féral 1989, Schröder 2003, Simo Bobda and Wolf 2003) and/or focus mainly on Cameroon Standard English (Mbangwana and Sala 2009, Wolf 2001).

We report on the construction of a 240,000-word pilot corpus of transcribed spoken CPE dialogues and monologues, with partial POS-tagging, glossing and translations. The proportions of monologue and dialogue are guided by the methodology of the International Corpus of English project, making our corpus immediately comparable with existing corpora of post-colonial varieties of English. The project is funded by a British Academy/Leverhulme small grant (ref. SG140663).

Besides operational and other expected difficulties in design (balance, representativeness) and compilation (collection, transcription, annotation), a corpus of a non-standard spoken variety poses certain challenges of its own. For example, despite its widespread use, CPE lacks a standard written form: an appropriate and properly motivated spelling system has to be developed prior to the transcription stage. Additionally, the intricacy of the language ecology in Cameroon makes identifying criteria for representativeness a challenge: although the project targets native speakers, there is considerable lectal variation as a consequence of the complex multilingual environment.

While a larger corpus is essential for investigating lexis, we illustrate how recurring grammatical patterns can still be investigated in a small corpus, despite the absence of POS-tagging. Even at a preliminary stage, 'raw' language data can (a) chart the distribution of certain known linguistic events, and (b) uncover evidence of new events.

We present a case study focusing on five high-frequency verbs in CPE, based on a small (100,000-word) 'pre-pilot' corpus of consisting of (i) spoken CPE (Ayafor, Green and Ozón,

in prep. (a)), (ii) existing published sources (Ayisi & Longinotto 2005; Bellama et al. 2006; Todd 1979), and (iii) elicited examples.

Focusing on the verbs *make*, *do*, *give*, *get* and *take*, we find evidence for a productive light verb strategy (Butt 2010) for the relexification of predicates (Wichmann and Wohlgemuth 2008), and observe that this small set of frequently occurring verbs participate both in light verb constructions (LVCs) (1) and in serial verb constructions (SVCs) (2). We also find evidence for (a) the grammaticalisation of *mek* 'make' as a marker of deontic modality, (b) a preference for the double object construction over the dative construction for *gif* 'give' ditransitives (contra Schröder 2013), and (c) the existence of benefactive *gif* 'give' SVCs in CPE (also contra Schröder 2013).

(1) no bi man di **mek** babisita, de wuman di **mek** babisita
'It's not the man who babysits, the woman babysits.'

(2) dem don kam lait lam **gif** wi
'They came and lit lamps for us.'

References

Ayafor, M., M. Green & G. Ozón. In prep (a). *A Spoken Corpus of Cameroon Pidgin English*. Ms, University of Sussex.

Ayisi, F. & K. Longinotto (dir.) 2005. *Sisters in Law* [DVD]. London: Vixen Films.

Bellama, D., S. Nkwelle & J. Yudom. 2006. *An Introduction to Cameroonian Pidgin*. 3rd ed. Cameroon Peace Corps.

Butt, M. 2010. The light verb jungle: still hacking away. In M. Amberber, B. Baker & M. Harvey eds. *Complex Predicates: Cross-Linguistic Perspectives on Event Structure*. Cambridge: CUP. 48-78.

Féral, C. de. 1989. *Pidgin-English du Cameroun. Description linguistique et sociolinguistique*. Paris: Peeters/Selaf.

Lewis, M. P., G. F. Simons & C. D. Fennig eds. 2014. *Ethnologue: Languages of the World*. 17th ed. Dallas, Texas: SIL International. http://www.ethnologue.com.

Mbangwana, P. N. & B. M. Sala. 2009. Cameroon English Morphology and Syntax*: Current Trends in Action*. Munich: LINCOM EUROPA.

Schröder, Anne. 2003. *Status, Functions and Prospects of Pidgin English: an Empirical Approach to Language Dynamics in Cameroon*. Tübingen: Gunter Narr Verlag.

Schröder, A. 2013. Cameroon Pidgin English structure dataset. In S. M. Michaelis, P. Maurer, M. Haspelmath, & M. Huber eds. *Atlas of Pidgin and Creole Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://apics-online.info/contributions/18. Accessed 2014-09-02.

Simo Bobda, A. & H.-G. Wolf. 2003. Pidgin English in Cameroon in the New Millenium. In P. Lucko, P. Lothar & H.-G. Wolf eds. *Studies in African Varieties of English*. Frankfurt: Peter Lang. 101-17.

Wichmann, S. & J. Wohlgemuth. 2008. Loan verbs in a typological perspective. In T. Stolz, D. Bakker & R. Salas Palomo eds. *Aspects of Language Contact. New Theoretical, Methodological and Empirical Findings with Special Focus on Romancisation Processes*, 89-121. Berlin/New York: Mouton de Gruyter.

Wolf, H.-G. 2001. *English in Cameroon*. Berlin: Mouton de Gruyter.

Todd, L. (editor & translator). 1979. *Some day been dey: West African Pidgin folktales*. London: Routledge & Kegan Paul.

Multilingual practices in Late Modern English: Baseline evidence from the *Corpus of Late Modern English Texts 3.0*

Päivi Pahta, Arja Nurmi, Jukka Tyrkkö & Jukka Tuominen (University of Tampere)

Multilingual practices, evidenced in the alternating use of resources from two or more languages by the same speaker/writer or by the interlocutors within one communicative episode, have recently attracted increasing interest in both present-day and historical linguistics. Recent research has established, for example, that multilingual practices are characteristic of language use in various types of English writings from different historical periods (see e.g. Schendl and Wright 2011). Although several genres and topic domains have received some attention in this body of research, most studies have been carried out on relatively small datasets allowing limited opportunities for generalization, and we are still lacking a credible overview of the frequency and types of switching practices. The increasing availability of large electronic corpora from different historical periods of English as well as new methods for identifying foreign content within and between passages of English makes it now possible to reach a more systematic overview of the frequency and types of multilingual practices through corpus-based study, and obtain baseline evidence of the phenomenon. Such baseline evidence is necessary for the analysis of multilingual practices of individuals or in individual texts in a particular period, or for identifying trajectories of change.

This paper presents a corpus-based analysis of multilingual practices in the Late Modern English period. The material consists of the 34-million-word multigenre *Corpus of Late Modern English Texts 3.0*, where the multilingual passages have been identified using a range of complementary automatic and semi-automatic techniques, including a new corpus tool, *Multilingualiser*, developed specifically for processing multilingual data. We have enhanced the corpus with sociolinguistic background information (e.g. gender, social status, age, education) regarding the authors. We have also expanded the basic text typological and bibliographic data assigned by the corpus compilers with variables such as the original format and the number of editions published in the first decade of publication, and possible triggering elements for multilingual passages, such as co-textual references to foreign locations (e.g. novels taking place in France vs. those confined to England). The enhanced data will allow us to present a fully evidence-based overview of (1) the overall frequency of foreign-language passages in written English in 1710–1920, (2) the variety of languages used in these texts in addition to English, (3) the breakdown of pragmatic types of foreign-language segments in different text types, and (4) the monofactorial and multifactorial analysis of multilingual practices with regard to the social and textual variables encoded.

References

Corpus of Late Modern English Texts 3.0 (CLMET3). Compiled by H. De Smet, H.-J. Diller & J. Tyrkkö. https://perswww.kuleuven.be/~u0044428/.

Schendl, H. & L. Wright eds. 2011. *Code-Switching in Early English*. Berlin/Boston: De Gruyter Mouton.

Negative concord in the language of British adults and teenagers

Ignacio M. Palacios Martínez (University of Santiago de Compostela)

Negative concord (NC) or double negation (e.g. *I don't know nobody in Havering, we don't never see him or talk about him*) has been studied quite extensively from a diachronic perspective (Nevalainen 1999, 2006; Rissanen 1999; Ulkaji 1999), focussing on its evolution and decline from Old English to the present. Jespersen (1917), Baker (1990) and Wouden (1997), among others, have also looked at different semantic and syntactic features of these negatives by distinguishing several types (resumptive, paratactic, litotes, etc) and examining their meanings in discourse. NC has also been discussed widely from a sociolinguistic perspective, mainly in terms of its pervasiveness and geographical differences in the most extended varieties of English, British English (Trudgill, 1990; Anderwald 2002, 2005) and American English (Wolfram & Schilling-Estes 2006), as well as in varieties such as African American Vernacular English (Labov 1972, Mazzon 2004), Bahamian English, Jamaican Creole and Australian Vernacular English (Kortman & Lukenheimer 2013).

This paper is intended to make a contribution to the existing sociolinguistic literature here by examining NC in the language of British adults and teenagers. I hypothesise that differences exist between these two groups of speakers, not only regarding frequency of use but also with respect to the patterns of NCs adopted. The pragmatics or the communicative intention of this type of negative will also be considered, especially in comparison with single negatives.

Preliminary findings, based on the analysis of data from three comparable adult and teenager corpora (COLT, BNC and Linguistic Innovators Corpus) indicate that: (i) NC is much more frequent in teenagers than in adults; (ii) the language of teenagers shows a wider variety of NC patterns than that of adults, with a richer system in the combination of negatives in NC clauses, and that this may be at least partly explained by the higher frequency of this phenomenon in the teen variety; (iii) although the number of negators which may combine with another negative in first position, and which also occupy a pre-verbal slot, is restricted to five negative words (*n't/not*, *never*, *nobody*, *nothing*, *hardly*), a total of eight different negative items are used as second negators in postverbal position (*nothing*, *no*, *no more*, *nobody*, *none*, *never*, *nowhere*, *neither*); (iv) NC with *not nothing*, *not no* and *not no more* seems to be the most common in both adult and teen productions; (v) the number of multiple negatives, that is, the presence of more than two negatives in the same clause (e.g. *I don't want nothing to do with you no more, no one haven't never said nothing*) is not as common as expected, and is restricted mainly to young speakers; (vi) the negative pronouns *nothing* and *nobody/no one* occupy initial position only very rarely in NC structures and tend to follow the negatives *n't/not* or *no*, which contrasts with the adverb *never*, which is recorded on 33 occasions preceding many different negatives; (vii) the number of resumptive negatives, that is, negative sentences followed by a supplementary negative outside the scope of the first one (Jespersen 1917), is very low, particularly those with *neither* and *nor* (e.g. *I don't like going to the pub and I don't drink neither*), and these are restricted to the language of teenagers; (viii) in this first analysis no significant pragmatic differences between the two groups considered are observed in the use of NC. This type of negative structure may be used by

speakers to accentuate a negative meaning, although this is not the norm, since in most cases NC structures are equivalent to single negatives.

References

Anderwald, L. 2002. *Negation in Non-standard British English. Gaps, Regularizations and Asymmetries*. London Routledge.

Anderwald, L.. 2005. Negative concord in British English dialects. In Y. Iyeri ed. *Aspects of English Negation*. Amsterdam: Benjamins, 116-37.

Baker, C. L. 1979. Double negatives. *Linguistic Inquiry* 1/2: 169-86.

Jespersen, O. 1917. *Negation in English and other Languages*. Copenhagen: Host.

Kortmann, B. & K. Lunkenheimer, eds. 2013. *The Electronic World Atlas of Varieties of English* [eWAVE]. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://www.ewave-atlas.org/, accessed on 2014-03-10.

Labov, W. 1972. Negative attraction and negative concord in English grammar. *Language* 48: 773–818.

Mazzon, G.. 2004. *A History of English Negation*. Harlow: Pearson Longman.

Nevalainen, T. 2006. Negative concord as an English "Vernacular Universal". Social history and linguistic typology. *Journal of English Linguistics* 34: 257-78.

Rissanen, M. 1999. Syntax. In R. Lass ed. *The Cambridge History of the English Language*, vol 3: 1476-1776. Cambridge: CUP, 187-331.

Singh, R. 1970. A note on multiple negatives. *American Speech*, 45: 247-51.

Trudgill, P. 1990. *The Dialects of England*. Oxford: Blackwell.

Ukaji, M. 1999. On the scope of negative concord. In Tieken-Boon van Ostade, I., G. Tottie, G. & W. van der Wurf eds. *Negation in the History of English*. Berlin & New York: Mouton de Gruyter.

Wolfram, W. & N. Schilling-Estes. 2005. *American English Dialects and Variation*. Malden/Oxford: Blackwell.

Wouden, T. van der. 1997. *Negative Contexts : Collocation, Polarity and Multiple Negation*. London: Routledge.

The impact of genre on EFL learner writing: A multi-dimensional analysis perspective

Magali Paquot (University of Louvain)
Douglas Biber (Northern Arizona University)

Previous learner corpus-based studies have shown that EFL learner languages exhibit shared linguistic features irrespective of the learners' first languages. For example, it has repeatedly been reported that EFL learner writing is characterized by a more involved style than the writing of their native peers, as evidenced by a high number of writer/reader (W/R) visibility features such as first and second person pronouns, *let's* imperatives, epistemic modal adverbs (e.g. *certainly*, *maybe*) and questions (cf. e.g. Petch-Tyson 1998, Altenberg and Tapper 1998; Aijmer, 2002; Narita & Sugiura, 2006; Neff et al. 2007; Gilquin & Paquot, 2008; Hasselgård, 2009).

Most of these learner corpus research (LCR) studies, however, have focused almost exclusively on argumentative writing and it is therefore questionable whether their results can be generalized to other genres and ultimately be used to inform English for Academic Purposes (EAP) pedagogical materials (Gilquin et al., 2007). To use the example of

involvement features again, as noted by Recski (2004), in the case of argumentative essays such as those contained in the widely used *International Corpus of Learner English* (ICLE, Granger et al. 2009), "personal references and subjective attitudes are certainly hard to avoid", since learners are explicitly prompted to give their personal opinions.

This study is part of a larger body of research that seeks to investigate whether the features commonly attributed to EFL learner writing are genuine characteristics of interlanguages or whether they are prompted by the argumentative type of texts that has usually been analysed in LCR. Paquot et al. (2013), for example, compared argumentative texts from the ICLE with discipline-specific texts from the *Varieties of English for Specific Purposes dAtabase* (VESPA). The VESPA learner corpus project aims at building a large corpus of English for Specific Purposes (ESP) texts written by L2 writers from various mother tongue backgrounds. The corpus currently contains papers and reports produced by BA and MA students in the context of a variety of content courses in linguistics, business, medicine, and engineering (for more details, see http://www.uclouvain.be/en-cecl-vespa.html). Paquot et al. (2013) analysed French and Norwegian learners' use of a variety of W/R visibility features in ICLE argumentative texts and VESPA papers written for linguistics courses. They showed that, when compared to native speakers' writing within the same genre and discipline, texts produced by French and Norwegian learners systematically displayed an overuse of W/R visibility features.

We adopt a broader perspective and build on a multi-dimensional analysis of a variety of native and learner corpora to further investigate the impact of genre on EFL learner writing. We make use of Biber's (1988) linguistic features and dimensions to compare and contrast the same ICLE and VESPA sub-corpora as used in Paquot et al. (2013) as well as two corpora of student native writing (comparable subsets of the *Louvain Corpus of Native English Essays* (LOCNESS), the *British Academic Written English* (BAWE) corpus) and a 1 million word corpus of published research articles in linguistics.

Preliminary results suggest that French and Norwegian learners' argumentative and discipline-specific texts are characterized by higher degrees of involvement (Dimension 1) and persuasiveness (Dimension 4) when set against comparable native speakers' writing in terms of genre and discipline. They also point to strong L1-based differences (e.g. Norwegian learners' argumentative and discipline-specific texts are much more involved than French learners' texts). However, the various corpora used also cluster by genre, irrespective of L1 background, thus suggesting that learners adapt to genre requirements to some extent (cf. Paquot et al., 2013).

The theoretical and pedagogical implications of this study will be discussed. We will also consider its implications in terms of corpus comparability and selection of a reference corpus in learner corpus research.

References

Aijmer, K. 2002. Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins, 55-76.

Altenberg, B. & M. Tapper. 1998. The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger ed. *Learner English on Computer*. London and New York: Addison-Wesley Longman, 80-93.

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

Gilquin, G., S. Granger & M. Paquot. 2007. Learner corpora: the missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4), 319-35.

Gilquin, G. & M. Paquot 2008. Too chatty: learner academic writing and register variation. *English Text Construction* 1(1): 41-61.

Granger, S., E. Dagneaux, F. Meunier, & M. Paquot. 2009 *The International Corpus of Learner English*. Version 2. Handbook and CD-Rom. Louvain-la-Neuve: Presses Universitaires de Louvain.

Hasselgård, H. 2009. Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In K. Aijmer ed. *Corpora and Language Teaching*. Amsterdam: Benjamins, 121-40.

Narita, M. & M. Sugiura 2006. The use of adverbial connectors in argumentative essays by Japanese EFL college students. *English Corpus Studies*, 13, 23-42.

Neff, J., F. Ballesteros, E. Dafouz, F. Martínez. & J.-P. Rica, 2007. A contrastive functional analysis of errors in Spanish EFL university writers' argumentative texts: corpus-based study. In E. Fitzpatrick ed. *Corpus Linguistics beyond the Word: Corpus Research from Phrase to Discourse.* Amsterdam: Rodopi, 203-25.

Paquot, M. 2010. *Academic Vocabulary in Learner Writing. From Extraction to Analysis*. London & New York: Continuum.

Paquot, M., H. Hasselgård, & S. Oksefjell Ebeling. 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. Proceedings of the First Learner Corpus Research Conference (LCR 2011).* Louvain-la-Neuve: Presses Universitaires de Louvain, 377-87.

Petch-Tyson, S. 1998. Reader/writer visibility in EFL persuasive writing. In S. Granger ed. *Learner English on Computer*. London & New York: Addison Wesley Longman, 107-18.

Recski, L.J. 2004. Expressing standpoints in EFL written discourse. *Revista Virtual de Estudos da Linguagem* 3. http://www.revel.inf.br/files/artigos/revel_3_expressing_standpoints_in_efl_written_discourse.pdf last accessed on December 15, 2014.

Indian English as a super-central variety: Diffusion of clause-final focus particles in Asian Englishes

Hanna Parviainen (University of Tampere)
Robert Fuchs (Münster University)

This study investigates the spread of clause-final focus particles from Indian English (IndE) to other Asian varieties of English. Previous research has found that the focus particles *also* and *only* have developed additional meanings in IndE, and they are often used clause-finally (Fuchs 2012, Lange 2007). Furthermore, Parviainen (2012) has shown that among Asian Englishes, this feature is most frequently used in IndE, but it also occurs in Philippine (PhiE)

and Hong Kong English (HKE). A possible explanation for this is that IndE has historically influenced other Asian Englishes. The present paper extends this research by using the apparent-time methodology to investigate ongoing language change in IndE, PhiE and HKE. Based on the respective components of the *International Corpus of English* (ICE), we show that female speakers in all varieties use clause-final *also* and *only* more frequently than male speakers. This tendency is strongest for clause-final *only*, whereas for clause-final *also*, the observed trend is weaker. We also find that younger speakers use clause-final *only* more often in all varieties, with the highest frequencies in IndE, followed by HKE and PhiE. For clause-final *also* we moreover find an age effect in HKE, but not in the other varieties. Based on this evidence, we conclude that clause-final *only* is a feature that is becoming more frequent in all varieties concerned. For *also*, the evidence is suggestive but not conclusive. Because the change appears to be most advanced in IndE, and less so in the other varieties, it is conceivable that IndE has influenced the other varieties; this has also been documented for other language features (Hundt et al. 2012), and is supported by historical ties and migration between these countries (Pluss 2006, Sridharan and Sevea 2006). Our results also support a description of IndE as a "super-central variety" (Mair 2013) that has considerable influence over other, geographically and culturally close varieties.

References

Fuchs, R. 2012. Focus marking and semantic transfer in Indian English. *English World-Wide,* 33(1): 27-52.

Hundt, M., S. Hoffmann, & J. Mukherjee. 2012. The hypothetical subjunctive in South Asian Englishes: local developments in the use of a global construction. *English World-Wide,* 33(2): 147-64.

Lange, C. 2007. Focus marking in Indian English. *English World-Wide,* 28(1): 89-118.

Mair, C. 2013. The world system of Englishes: accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide,* 34(3): 253-78.

Parviainen, H. 2012. Focus particles in Indian English and other varieties. *World Englishes,* 31(2): 226-47.

Pluss, C. 2006. Hong Kong. In B. V. Lal, P. Reeves & R. Rai eds. *The Encyclopedia of the Indian Diaspora.* Honolulu: University of Hawai'i Press, 206-09.

Sridharan, K. & T.S. Sevea. 2006 The Philippines. In B. V. Lal, P. Reeves & R. Rai eds. *The Encyclopedia of the Indian Diaspora.* Honolulu: University of Hawai'i Press, 198-99.

The variability of strong verb past forms in C21 Englishes

Pam Peters (Macquarie University, NSW Australia)

Steady reduction in the forms used in the declination of English strong verbs is visible over the course of centuries. Yet the various factors that contribute to their deconstruction are not obvious in any synchronic sample of language. The variability in strong verb past forms is more evident in some regional varieties than others, as can be seen in comparative data from British, Australian and New Zealand ICE corpora (Peters, 2009). Change might be predicted to occur more readily in the high-contact rather than the lower-contact varieties of English (Kortmann 2014); and even more so in indigenized varieties where the raw material of

English is reconstructed in a multilingual habitat. The evolutionary stage reached by each variety, in terms of exonormativity or endonormativity (Schneider 2007), makes it less or more responsive to internal change.

The pace of language change has also been found to correlate with the relative frequency of the variable. Regularization occurs more rapidly with lower frequency, less visible items, and is resisted by higher frequency ones despite their irregularities, as underscored in Lieberman et al.'s statistical study (2007) of the regularization of English verbs over the centuries. But this has yet to be complemented by large-scale synchronic data on variation within individual verb classes, so as to identify those most susceptible to variation and regularization. The compounding factors of token and type frequency (Bybee & Thompson 1997) are also to be taken into account. Synchronic data could be used to update the conventional strong verb classes, instead of relying on historical or morphophonological criteria.

Ample data on English strong verb usage is now available through GloWbE, a very large (1.9 billion-word) corpus of twenty varieties of English (Davies & Fuchs 2015), including low- and high-contact varieties and indigenised varieties. Because the GloWbE data was compiled from websites and blogs in 2012, it is consistent in time, and relatively free of the constraining processes of editing and print publishing. This study focuses on variation in the past forms of the verbs *begin, drink, ring, shrink, sing, sink, spring, stink, swim* , i.e. class 7A in Quirk et al.'s (1985) classification. Data extracted from six GloWbE varieties of English are examined (British, American, Canadian, Australian, Singaporean, and from the Philippines), representing a range from low to high contact varieties.

In the corpus data analysed so far, the patterns of variation show only limited regional effects correlating with their relative levels of varietal contact or endonormativity. But two kinds of frequency effects are evident: (i) the lowest frequency verbs in the set (*spring, stink)* make greater use of the "u" (past participle form) for the past tense; whereas (ii) the high frequency verbs *(begin, drink, sing)* show variable use of the "a"(past tense form) for the past participle. Both alternative patterns have the vowel contrasts reduced from three to two within the paradigm, and aligned respectively with Class 6B (*fling, slink)* and Class 6E (*sit, spit)*. Within Class 6 there are also signs of variability, in scattered instances of past forms with *–ed* for both low- and higher frequency verbs (e.g. *digged, slinked).* The underlying tendency for strong verbs to regularize – to reduce to two contrasting forms, whatever their relative frequency – is reflected in all this variation.

References

Bybee, J. & S. Thompson. 1997. Three frequency effects in syntax. *Annual Proceedings of the Berkeley Linguistics Society.* Online eLanguage.

Davies, M. & R. Fuchs. 2015. Expanding Horizons in the study of world Englishes with the 1.9 Billion Word Global Web-Based English Corpus GloWbE. *English World-Wide,* 36:1.

Kortmann, B. 2014. Looking up how grammars vary across the English-speaking world: the Wave perspective. Plenary at AAH Symposium, Canberra 20-21 Nov 2014.

Lieberman, E. et al. 2007. Quantifying the evolutionary dynamics of language. *Nature,* 449 October, 713-16.

Peters, P. 2009. Irregular verbs: regularization and ongoing variability. In P. Peters et al eds. *Comparative Studies in Australian and New Zealand English.* Amsterdam: Benjamins.
Quirk, R. et al 1985. *Comprehensive Grammar of the English Language.* London: Longman.
Schneider, E. 2007. *Postcolonial English.* Cambridge: Cambridge University Press.

The peripheral-specific meanings of epistemic-evidential complement-taking predicates in English

Nele Põldvere & Carita Paradis (Lund University)

Epistemic and evidential *complement-taking predicates* (CTPs) are constructions that may be used in various positions in an utterance. The following examples exemplify their use in initial (1), medial (2) and final positions (3) relative to the proposition they modify:

(1) *I suppose* this is the complete choice.
(2) I myself would never *I think* expect a verbal statement worked out at a first meeting.
(3) He's working for a PhD here *I think*.

The main objective of the present study is to find out if CTPs are also able to indicate peripheral-specific meanings. The idea of peripheral-specific meanings is a relatively new development in linguistic research, despite the long history of the phenomenon itself (Traugott, 2012; Degand, 2014). This is most likely due to the relative infrequency of discourse markers at right periphery compared to the left periphery in English (Traugott, 2013). The phenomenon entails that discourse markers are associated with different meanings relative to the position they occupy in an utterance. According to this view, the left periphery is typically associated with speaker-oriented, subjective meanings, while the right periphery attracts addressee-oriented, intersubjective meanings. We adopt Traugott's (2010) definitions of subjectivity and intersubjectivity, in that the former refers to the speaker's awareness of his/her own attitudes and viewpoints, while in case of the latter, the awareness is directed at the addressee's self-image. Although medial positions do not constitute clause periphery, the study will also try to uncover whether medial positions behave more similarly to the left or right periphery.

The data come from the London-Lund Corpus of spoken British English (Svartvik & Quirk, 1980), and more specifically, from face-to-face spontaneous dialogues between educated adults. This allows us to study interaction in its most natural form where parenthetical CTPs are most likely to occur. Another advantage of the corpus is its close and detailed prosodic annotation, since prosody has been shown to be an important indicator of the semantic-pragmatic features of CTPs (Dehé & Wichmann, 2010). The parameters to be explored include positional, prosodic and functional factors that are believed to either confirm or refute the existence of peripheral-specific meanings of CTPs. Couched in the framework of Construction Grammar (Goldberg, 1995), the constructions are viewed as part of a larger constructional family whose members are synonymous with regards to their association with epistemic-evidential meanings. The methods chosen for the study are both qualitative and

quantitative in order to be able to give a comprehensive overview of the use of these constructions in context.

Our initial results indicate that when it comes to CTPs then the general tendencies associated with peripheral-specific meanings can only be observed to a certain degree. Instead, CTPs seem to display intersubjectivity in all positions in the utterance, although these functions exemplify intersubjectivity differently. While the left periphery often acts as a site for expressions related to face-saving and politeness, the prosodic cues of CTPs at the right periphery imply that these constructions regularly facilitate turn-taking on the part of the addressee. Also, although synonymous in their use as epistemic-evidential markers, the functions of these constructions are shown to be dependent on the type of predicate chosen, with more frequent constructions being the most salient representatives of the tendencies noted above.

References

Degand, L. 2014. 'So very fast very fast then'. Discourse markers at left and right periphery in spoken French. In K. Beeching & U. Detges eds. *Discourse Functions at the Left and Right Periphery: Crosslinguistic Investigations of Language Use and Language Change*. Leiden: Brill, 151-78.

Dehé, N. & Wichmann, A. 2010. The multifunctionality of epistemic parentheticals in discourse: prosodic cues to the semantic-pragmatic boundary. *Functions of Language*, 17 (1): 1-28.

Goldberg, A. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Svartvik, J. & R. Quirk eds. 1980. *A Corpus of English Conversation*. Lund: Lund Studies in English 56.

Traugott, E. C. 2010. Revisiting subjectification and intersubjectification. 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.

Traugott, E. C. 2013. *I must wait on myself, must I? On the rise of pragmatic markers at right periphery of the clause in English*. Paper presented at a meeting at Lund University, Sweden, September 2013.

Diachronic cross-genre comparisons in the use of phrasal verbs (1650-1990)

Paula Rodríguez Puente (University of Cantabria)

The term phrasal verb or verb particle combination refers to a sub-type of multi-word verbs consisting of a verb plus a particle of adverbial nature (cf. Claridge 2000: 46) which functions as a single unit to varying degrees. Present-day English phrasal verbs tend to be associated with spoken, colloquial registers (Biber et al. 1999: 408, 409). Based on evidence from the *Corpus of Nineteenth-Century English*, the studies by Kytö & Smitterberg (2006) and Smitterberg (2008) have concluded that by the 19th century the frequency of phrasal verbs also correlates with the degree of formality of the text. Some studies point out that this was also the case during the Early Modern English period (cf., e.g., Hiltunen 1994;

Nevalainen 1999: 423; Claridge 2000: 185-197). This statement has been lately challenged by Thim (2006, 2012), who argues that in Early Modern English the (non)-occurrence of phrasal verbs in a particular text seems rather motivated by its contents, which may prompt the use of phrasal verbs to convey literal meanings predominantly, whereas the degree of formality is a secondary aspect. Based on evidence from the British section of ARCHER (*A Representative Corpus of Historical English Registers*, 1650-1990) and the *Old Bailey Corpus* (1720-1913), this paper compares the use of phrasal verbs in written register (fiction, medicine, news, science), registers which show different degrees of spechlikeness (diaries, drama, journals, letters, sermons), and those which (intend to) reproduce the spoken language of the past (trial proceedings). As argued by Huber (2007), the *Proceedings of the Old Bailey* contain verbatim passages which "are arguably as near as we can get to the spoken word of the period," thus offering the opportunity to analyse the everyday language of the past. Previous research has shown that the evolution of phrasal verbs in different text types follows the stylistic drift described in the multivariate analyses by Biber and associates (see, among others, Biber 1988, Biber & Finegan 1989, 1997). However, other factors such as the topics dealt with in the text or the format chosen to portray the narration may also affect the use and frequency of phrasal verbs. The wide variety of text types analysed in this paper facilitates comparisons among the syntactic, semantic and morphological characteristics of phrasal verbs. Special attention will be paid to any differences or similarities in the possible syntactic arrangements of the past as compared to PDE, the behavior of semantically transparent as opposed to idiomatic combinations, as well as the divergence in verb and particle types used across genres. The data also yield more detailed knowledge of the development of the use of phrasal verbs, not only between oral vs. written texts but also between formal and informal registers. This presentation will offer evidence that the stylistic and formal features of phrasal verbs differ across registers, not only in PDE but also in previous stages of the language.

References

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

Biber, D. & E. Finegan. 1989. Drift and the evolution of English style: a history of three genres. *Language*, 65 (3): 487-517.

Biber, D. & E. Finegan. 1997. Diachronic relations among speech-based and written registers in English. In T. Nevalainen & L. Kahlas-Tarkka eds. *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen.* Helsinki: Société Néophilologique; 253-75.

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

Claridge, C. 2000. *Multi-Word Verbs in Early Modern English: A Corpus-Based Study.* Amsterdam: Rodopi.

Hiltunen, R. 1994. Phrasal verbs in Early Modern English: notes on lexis and style. In D. Kastovsky ed. *Studies in Early Modern English*. Berlin & New York: Mouton de Gruyter, 129-40.

Huber, M. 2007. The *Old Bailey Proceedings*, 1674-1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English. *Studies in Variation, Contacts and Change in English 1: Annotating Variation and Change*. http://www.helsinki.fi/varieng/series/volumes/01/huber/

Kytö, M. & E. Smitterberg. 2006. Nineteenth-century English: an age of stability or a period of change? In R. Facchineti & Matti Rissanen eds. *Corpus-Based Studies of Diachronic English*. Bern: Peter Lang; 199-230.

Nevalainen, T. 1999. Early Modern English lexis and semantics. In R. Lass ed. *The Cambridge History of the English Language: Vol 3: 1476-1776*. Cambridge: Cambridge University Press, 332-458.
Smitterberg, E. 2008. The progressive and phrasal verbs: Evidence of colloquialization in nineteenth-century English? In T. Nevalainen, I. Taavitsainen, P. Pahta & M. Korhonen eds. *The Dynamics of Linguistic Variation. Corpus Evidence on English Past and Present* (*Studies in Language Variation* 2). Amsterdam: Benjamins; 269-89.
Thim, S. 2006. Phrasal verbs in *Everyday English*: 1500-1700. In A. J. Johnston, F. von Mengden & S. Thim eds. *Language and Text: Current Perspectives on English and Germanic Historical Linguistics and Philology*. Heidelberg: Winter, 291-306.
Thim, S. 2012. *Phrasal Verbs. The English Verb-Particle Construction and its History* (*Topics in English Linguistics* 78). Berlin and New York: Mouton de Gruyter.

A frequency-based approach to salience in dialect grammar: The case of Welsh English

Katja Roller (Research Training Group DFG 1624 "Frequency Effects in Language", University of Freiburg)

English in Wales features some grammatical constructions that show clear influences from the Welsh substrate (cf. Paulasto 2006) and, thus, differ from other varieties of English in the British Isles. But to what extent are people inside and outside of Wales aware of those differences? Are there some grammatical features that stick out, that are more *salient* than others? Moreover, to what extent are the features' saliences connected to their frequencies in language use?

Linguistic salience is defined by Kerswill and Williams (2000) as "a property of a linguistic item [...] that makes it in some way perceptually and cognitively prominent". Several sociolinguistic and dialectological works assume frequency to be a crucial determinant of salience (cf. e.g. Rácz 2013, Labov et al. 2011, Auer 2014), probably interacting with structural-linguistic and attitudinal factors (cf. e.g. Trudgill 1986, Kerswill & Williams 2000). However, so far there is no study systematically analysing the interplay of usage frequencies and salience in dialect grammar. This paper explores to what extent the salience of a morphosyntactic dialectal feature can be predicted by its frequency of use. I hypothesise that highly salient features are frequent in contrast to (1) other features in the target variety (Welsh English), (2) uses of the same feature in (an)other language variety/ies (e.g. London English) (cf. Kerswill & Williams 2002, Rácz 2013). By implementing a quantitative method, I aim to make the complex concept of salience more comprehensible and easier to approach.

The paper is roughly divided into three parts. First, it is determined which features of Welsh English grammar are salient – both to people from Wales (intralectal salience) and to people from London (interlectal salience). The data come from a questionnaire-based survey in Wales and London (2013-2014, 300 subjects). Second, I investigate to which degree salience can be attributed to frequency. The frequencies of grammatical features are determined via corpus analyses. For Welsh English, I use my self-transcribed Radio Wales Corpus (270,000

words, interviews from 1999-2005). This corpus with spoken data is employed since the grammatical features investigated in my study are reported to mainly occur in spoken conversation (cf. Trudgill & Hannah 2002). To find out about the frequencies of the features in London English, the Linguistic Innovators corpus is consulted (1.1 million words, 2004-2005, cf. Kerswill et al. 2007). Third, this talk explores the impact of other potential causes of salience (e.g. language attitudes). The data here also come from the 2013 to 2014 questionnaire-based survey.

First results suggest that particularly salient features consciously associated with Welsh English are focus fronting (*A student he was*, cf. Paulasto 2006) and the invariant tag question *isn't it* (*You like him, isn't it?*). In the salience survey, these constructions received significantly higher ($p<0.001$) salience values than the other features under investigation, by both the Welsh and the London informants. As for the determinants of salience, the corpus analyses suggest a positive relation between absolute frequencies in Welsh English and salience in that the more frequent features are generally more salient. However, the correlation is only strong for intralectal salience, i.e. the salience to people from Wales ($r=0.57$), while it is intermediate for interlectal salience ($r=0.39$). It seems that a more powerful predictor of the salience to people from London are relative frequency differences between Welsh English and London English, which correlate with salience at $r=0.72$. Furthermore, my findings suggest that language attitudes and individual characteristics of the informants also have an influence on salience.

By interrelating results from corpus linguistic, perceptual and attitudinal studies, this project offers a novel approach to linguistic salience. More generally, the study aims to offer empirical evidence for usage-based theories of language (cf. e.g. Bybee 2006) by showing that salience is grounded in frequencies of occurrence in language use.

References

Auer, P. 2014. Anmerkungen zum Salienzbegriff in der Soziolinguistik. *Linguistik Online*, 66.4.

Bybee, J. 2006. From usage to grammar: the mind's response to repetition. *Language*, 82:4: 711-33.

Kerswill, P. et al. 2007. *Linguistic Innovators: the English of Adolescents in London. Full Research Report*. ESRC End of Award Report, RES-000-23-0680. Swindon: ESRC.

Kerswill, P. & A. Williams. 2000. "Salience" as an explanatory factor in language change: evidence from dialect levelling in urban England. *Reading Working Papers in Linguistics* 4, 63-94.

Kerswill, P. & Williams, A. 2002. Dialect recognition and speech community focusing in new and old towns in England: the effects of dialect levelling, demography and social networks. In D. Long & D. Preston eds. *A Handbook of Perceptual Dialectology*. Vol 2. Amsterdam: Benjamins, 178-207.

Labov, W. et al. 2011. Properties of the sociolinguistic monitor. *Journal of Sociolinguistics*, 15.4: 431-63.

Paulasto, H. 2006. *Welsh English syntax: Contact and variation*. Joensuu: Joensuu University Press.

Rácz, P. 2013. *Salience in Sociolinguistics: A Quantitative Approach*. Berlin and Boston: Mouton de Gruyter.

Trudgill, P. 1986. *Dialects in Contact*. 1st ed. Oxford: Blackwell.

Trudgill, P. & J. Hannah. 2002. *International English: A Guide to the Varieties of Standard English*. 4th ed. London: Arnold.

How does constructional knowledge emerge? Evidence from a longitudinal corpus of German and Spanish learner English

Ute Römer (Georgia State University)
Cynthia M. Berger (Georgia State University)
Nick C. Ellis (University of Michigan)

This paper takes a usage-based, emergentist perspective (Ellis & Larsen-Freeman, 2006) on language acquisition to investigate how the use of verb-argument constructions (VACs) develops in the writing of second language learners across proficiency levels. We present findings from an analysis of L1 German and L1 Spanish learner use of English VACs, such as the 'V *about* n' (e.g., *she thinks about chocolate a lot*) or the 'V *in* n' construction (e.g., *he lives in a small town*). Our analysis is based on a corpus of learner writing at different levels of proficiency, described in further detail below. We were interested in determining (1) how VACs develop in second language (L2) writing as proficiency increases and (2) how the use and emergence of VACs is affected by the learner's first language.

The paper builds on previous work on learner knowledge of VACs (Römer et al., 2014a and 2014b; Gries & Wulff, 2005) that has shown that advanced learners of English have constructional knowledge, that learners' VAC knowledge differs in systematic ways from that of native speakers, and that learners' verb-VAC associations differ across L1 groups. What previous studies have not been able to address, mostly due to the unavailability of pertinent data at lower proficiency levels, is how this constructional knowledge unfolds over time. Likewise, only few studies have systematically contrasted learners from different L1 backgrounds to investigate the role of transfer from the first language in more depth. The present paper seeks to take first steps to closing both of these gaps.

Large amounts of texts produced by L2 learners representing a range of L1 backgrounds and proficiency levels have recently been made available by the Education First (EF) research unit at the University of Cambridge, UK in the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen, Alexopoulou, & Korhonen, 2013). For our study, we retrieved sets of texts written by German and Spanish learners at Common European Framework of Reference (CEFR) levels A1 through C2 from EFCAMDAT. The EFCAMDAT subsets we compiled—over 28,000 texts and 2.8 million words from L1 German learners, and over 40,000 texts and 3.2 million words from L1 Spanish learners—constitute a quasi-longitudinal learner corpus that complements existing corpus resources. Using prepositions as seed words, we extracted over 20,000 tokens of 19 different VACs from the EFCAMDAT subsets.

The EFCAMDAT datasets enabled us to describe learners' dynamically evolving abilities and trace the emergence of constructions over time. We will report trends in frequency developments of VACs, type/token ratios, dominant verb-VAC associations, semantic developments of verbs in VACs, correlations between verbs produced by learners at different levels, and correlations between verbs produced by learners of different first languages (German vs. Spanish). In our comparison of German and Spanish learner production of VACs across levels, we considered issues of language transfer and typology that affect the verb system, particularly verb semantics, and make reference to Talmy's (1985) distinction

between verb-framed and satellite-framed languages. We found an increase in the range of VACs and their productivity from A1 to C2 levels. Our findings also suggest that the most strongly associated verbs in learner VACs change with proficiency from denoting general to more specific actions, which correlates in turn with increasingly higher shares of verbs from lower-frequency bands at higher proficiency levels. Observed marked differences between German and Spanish learner development of VACs can largely be explained on the basis of language typology and transfer.

References

Römer, U., M. B. O'Donnell & N. C. Ellis. 2014a. Second language learner knowledge of verb-argument constructions: effects of language transfer and typology. *The Modern Language Journal*, 98(4): 952-75.

Römer, U., A. Roberson, M. B. O'Donnell & N. C. Ellis. 2014b. Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal*, 38: 59-79.

Ellis, N. C., & D. Larsen-Freeman. 2006. Language emergence: implications for applied linguistics. Introduction to the special issue. *Applied Linguistics, 27*(4): 558-89.

Geertzen, J., T. Alexopoulou & A. Korhonen. 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.

Gries, S. T., & S. Wulff. 2005. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics*, 3: 182-200.

Talmy, L. 1985. Lexicalization patterns: semantic structure in lexical form. In T. Shopen ed. *Language Typology and Syntactic Description: Grammatical Categories and the Lexicon.* Cambridge: Cambridge University Press, 57-149.

Turn order and turn distribution in storytelling

Christoph Rühlemann (University of Paderborn)
Stefan Th. Gries (University of California, Santa Barbara)

In this paper, we aim to contribute to the growing field of corpus pragmatics (cf. Aijmer & Rühlemann, forthcoming) focusing on turntaking patterns in conversational storytelling. It has long been noted that turntaking in every-day narrative differs on a number of counts from turntaking in regular conversation. The differences, however, have, at best, been researched qualitatively based on casual observations and small datasets. Here, we base our analysis on large data sets extracted from the British National Corpus (BNC) as well as two specialized corpora of conversational narrative, the SCoSe (Saarbrücken Corpus of Spoken English) containing American English 4- and 5-party stories and the NC (Narrative Corpus) containing British English 4- to 7-party narratives from the BNC. The analysis is decidedly quantitative and statistical in orientation. Specifically, we are concerned with *turn order* in conversational multi-party narrative aiming to examine the validity of Sacks' description of storytelling as "an attempt to control a third slot in talk, from a first" (Sacks 1992: 18), a turn order pattern referred to as the N-notN-N pattern. We also investigate whether individual turntaking styles have an impact on *turn distribution* (a measure intimately related to turn order). Further, given the structural differences in the data at hand (the SCoSE being raw-

text, the NC being densely annotated) we employ largely different methodologies particularly in addressing our main question, which is related to turn order.

The results on turntaking style suggest that this factor cannot on its own account for the noticeable increase in the narrator's turn share as soon as the conversational activity moves into storytelling. The results on turn order reveal the N-notN-N pattern's statistical overrepresentation in all multi-party narrative types examined. The implications of this finding are far-reaching. First, Sacks et al.'s (1974) dictum that turn order is not fixed in advance does not hold true for conversational narrative. Also, turn order in conversational narrative is not locally controlled on a turn-by-turn basis, but globally on the basis of the activity the conversationalists are involved in, viz. storytelling.

Second, a fundamental correlate of the N-notN-N pattern is the avoidance of double-responses, that is, of two consecutive response turns following the narrator's turn. This avoidance suggests that the turn order system underlying multi-party narrative is that of 2-party talk. Further, the double-response avoidance suggests the possibility that the source of the turn-order bias in narrative is a tacit agreement between the recipients to promote the recipient filling the single-response slot to a 'spokesperson' taking the turn *on behalf* of all other recipients. We also note the possibility of there being a recipient-subsystem for turntaking at the single-response slot (the notN in the N-notN-N pattern) interacting with the narrator-recipient turntaking organization but still, to an extent, working on its own terms.

References

Aijmer, K. & C. Rühlemann. Forthcoming. *Corpus Pragmatics: A Handbook.* Cambridge: Cambridge University Press.

Sacks, H. 1992. *Lectures on Conversation*. Vols 1 & 2. Oxford: Blackwell.

Sacks, H., E. A. Schegloff & G. A. Jefferson. 1974. A simplest systematics for the organization of turn-taking in conversation. *Language*, 50(4): 696-735.

The influence of corpus-based frequency features on English compound spelling

Christina Sanchez-Stockhammer (Erlangen-Nuremberg)

Compound spelling variant selection is a ubiquitous lexis-related phenomenon: whenever language users write a compound in English, they need to select some way of linking the constituents orthographically. The three most common spelling variants for biconstituent compounds are open spelling with a space between the constituents (*drinking fountain*), hyphenation (*far-reaching*) and concatenated solid spelling (*teamwork*).

While the view that this phenomenon is completely unsystematic can be frequently found in the literature (e.g. Merriam-Webster 2001), my recent large-scale empirical study shows that underlying patterns can actually be observed and that corpus-based frequency measures play a significant role in English compound spelling. The study goes beyond previous research on the spelling of English compounds (e.g. Sepp 2006; Rakić 2009; Kuperman and Bertram

2013) by not limiting itself to noun+noun combinations. It investigates the influence of about 40 features (e.g. length, part of speech, stress pattern) on a list of over 10,000 compounds and focuses on 600 biconstituent compounds with highly established spelling variants in British English (200 each with uniquely open, hyphenated and solid spelling in five to six dictionaries from various publishers).

A substantial proportion of the features considered in the study is frequency-related and was therefore investigated by means of corpora. For instance, the written component of the British National Corpus (BNCwritten) was used in order to determine the effect of the following frequency-based features on compound spelling:

- the frequency of the whole compound
- the part-of-speech-sensitive frequency of the individual constituents in a lemmatised frequency list of BNCwritten (extracted from the CQP Edition of BNCweb at http://bncweb.lancs.ac.uk)
- the ratio between the frequencies of the constituents in the lemmatised frequency list
- the spelling-sensitive frequency of the compound in attributive position (directly preceding a noun) and in predicative position (directly following a verb)
- the spelling-sensitive left and right constituent family frequency, i.e. the summed frequencies for all compounds sharing the left and right constituent with the search item, respectively (cf. Plag 2010: 244). The underlying constituent family was extracted automatically from the digital list of entries in the Macmillan English Dictionary for Advanced Learners.

The BNCwritten was used, since it represents a comparatively large and accessible balanced corpus of relatively recent British English. In addition, even more recent language use was covered by the consideration of the one-million-word BE06 corpus (comprising texts from the years 2003-2008) and corpora of newer registers such as text messages (CorTxt Corpus), blogs (Blog Authorship Corpus) and chat communication (NPS Chat Corpus). The results of the statistical testing, which will be presented in detail in the paper, revealed a significant effect of a large number of corpus-based independent variables on the dependent variable 'compound spelling'. The paper concludes with a discussion of the role of the frequency-based corpus-derived features as determinants of English compound spelling in comparison to the other types of feature investigated.

References

Kuperman, V. & R. Bertram. 2013. Moving spaces: Spelling alternation in English noun-noun compounds. *Language and Cognitive Processes* 28(7). 939-966.

Macmillan English dictionary for advanced learners (MED). 2007. 2nd edn. with CD-ROM. Oxford: Macmillan.

Merriam-Webster. 2001. *Merriam-Webster's guide to punctuation and style*. 2nd edn. Springfield, MA: Merriam-Webster.

Plag, I. 2010. Compound stress assignment by analogy: The constituent family bias. *Zeitschrift für Sprachwissenschaft* 29(2). 243-282.

Rakić, S. 2009. Some observations on the structure, type frequencies and spelling of English compounds. *SKASE Journal of Theoretical Linguistics* 6. http://www.skase.sk/Volumes/JTL13/pdf_doc/04.pdf. (26 April, 2012.)

Sepp, Mary. 2006. *Phonological Constraints and Free Variation in Compounding: A Corpus Study of English and Estonian Noun Compounds*. The City University of New York: PhD dissertation.

New perspectives on gathering, vetting and employing Big Data from online social media: An interdisciplinary approach

Teri Schamp-Bjerede (Lund University)
Carita Paradis (Lund University)
Kostiantyn Kucher (Linnaeus University)
Andreas Kerren (Linnaeus University)
Magnus Sahlgren (Gavagai AB, Stockholm)

Massive textual data sets available in online social media are valuable resources for research in linguistics compared to static corpora in that they offer the possibility of analyzing the temporal unfolding of communication in addition to a wide variety of forms, styles and topics. By analyzing such data, linguists are able to investigate the ongoing evolution of language use in text and discourse. There are however also disadvantages with big data sets of this kind, such as problems of navigation and selection of sample texts, lack of annotation, abundance of neologisms and ungrammatical expressions. In order to be able to make sensible and optimal use of these data sets, the analysis requires advanced computational methods—both automated and semi-automated (from basic concordances and to more complex natural language processing)—accompanied by visualization techniques that are needed both for the visualization itself and for the interpretation of the results.

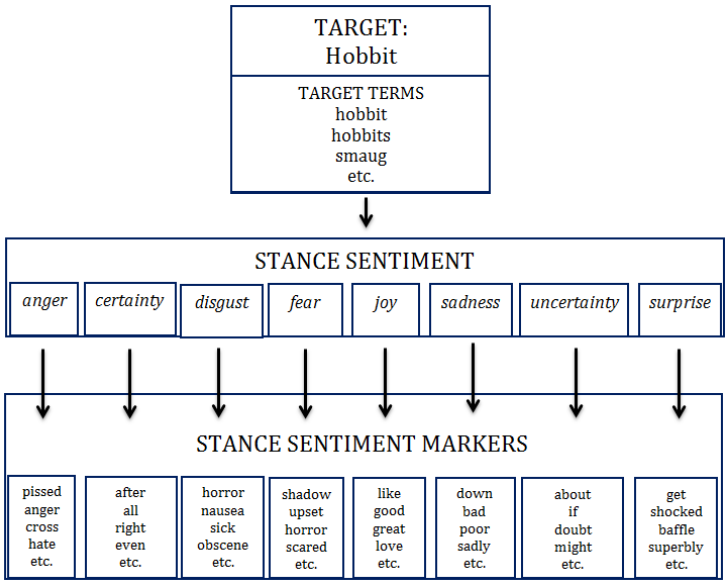


Figure 1 uVSAT logic map

The purpose of this paper is to present a use case to demonstrate our approach to Big Data research using insights from linguistics, natural language processing, and information visualization/visual analytics (Thomas & Cook 2006). The use case is concerned with the analysis of social media texts related to the movie series *The Hobbit*. The data selected for

analysis consists of microblog and forum posts before and after the release of *The Hobbit: The Desolation of Smaug* in December 2013.

At this point in time, our visual analytics tool uVSAT (Kucher et al. 2014) is designed to use a sentiment analysis approach (Figure 1). We use the sentiment markers (words), expressing six core emotions (Ekman 1992) and two epistemic stance categories, CERTAINTY / UNCERTAINTY, to identify the interlocutors' degree of commitment and attitudes to what they are discussing and in relation to the sentiments expressed. Through the use of such markers, we are able to identify and later analyze when heated discussions occur, how they occur in the flow of communication and in relation to what events in the world, as well as in the discussion itself among the participants. This will give us an idea about exactly what words and constructions are used to express the different sentiments and stances towards what is talked about, when, and where, and by whom (Du Bois 2007). That information can then be used to interact with the data in various ways. By using our visual analytics tool, we are able to analyze the temporal developments in social media, fetch the sets of relevant HTML documents, analyze the distribution of markers in the retrieved data, explore the text content with assistance of multiple interactive techniques, and export new markers as well as processed documents (Figure 2). The temporal trends as well as the distribution of sentiment and stance markers pertaining to the various categories, anger, joy, surprise, or certainty, etc., can be used to make predictions about sentiment and stance expressions in social media that will occur with regard to, in this particular use case, the subsequent movie "The Hobbit: The Battle of the Five Armies". After the premiere in December 2014, we will analyze the actual text data and compare the outcome of that with our predictions.

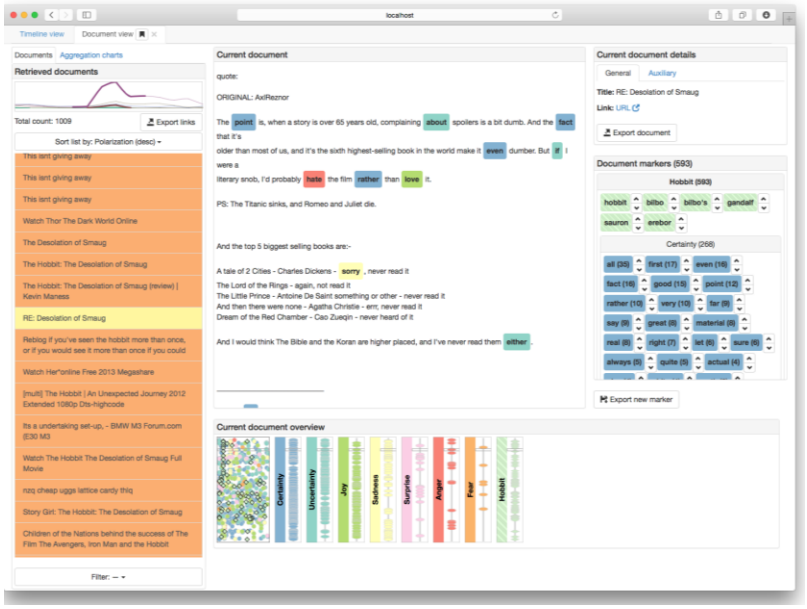


Figure 2 Document view

The contributions of this work to the corpus community include: (i) the introduction of a multidisciplinary approach to sentiment and stance analysis involving social media text data

instead of static corpora, (ii) the description of the combination of visualization and interaction techniques that facilitate the linguistic analysis of textual data, and (iii) a use case demonstrating how insights can be gained from large amounts of social media texts through these techniques.

References

Du Bois, J.W. 2007. The stance triangle. In R. Englebretson eds. *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam: Benjamins, 139-82.

Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4): 169-200.

Kucher, K., A. Kerren, C. Paradis, & M. Sahlgren. 2014, November. Visual analysis of stance markers in online social media. Poster session presented at IEEE Visual Analytics Science and Technology (VAST'14), Paris, France.

Thomas, J. J., & K.A. Cook. 2006. A visual analytics agenda. *Computer Graphics and Applications, IEEE*, 26(1): 10-13.

Why should *happy* people be *glad* and *closed* eyes be *shut*? Synonym selection as a strategy of stress clash avoidance

Julia Schlüter & Gabriele Knappe (University of Bamberg)

There seems to be agreement in scholarship that "absolute synonymy is rare—and when found mostly fleeting" (Dolezal 2013: 255). Clearly, absolute synonymy is avoided for reasons of language economy. It has been shown that transient situations of absolute synonymy tend to give way to splits, resulting either in semantic differentiation or stylistic divergence (cf., e.g., Cruse 2002: 489).

This paper argues that (near) synonymy can evolve into a third type of split, viz. a syntactic specialization, and it tests the hypothesis that one major factor underlying such fixation may be rhythmic in nature. Thus, the analysis takes the study of the preference for alternating stressed and unstressed syllables, which is receiving an increasing amount of attention in research on grammatical variation phenomena (e.g. Schlüter 2005; Ehret, Wolk & Szmrecsanyi 2014; Shih, Grafmiller, Futrell & Bresnan to appear), one level further, extending it to lexical choices.

It has been anecdotally noted that the selection of (close) synonyms can be sensitive to the avoidance of adjacent stressed syllables: Based on Bolinger's (1965: 149) observation that ?*a glád dáy* seems objectionable, while *a glád occásion* or *a háppy dáy* are acceptable, we investigate the syntactic distribution of (near-)synonymous adjective pairs such as *glad* vs. *happy*, *rich* vs. *wealthy*, *quick/fast* vs. *rapid*, etc. showing that the monosyllabic members tend to be underrepresented in attributive position. The hypothesis being pursued is that, due to the pervasiveness of initial stress in English nouns, monosyllabic adjectives – while unproblematic in predicative uses – tend to be avoided in attributive position if they can be replaced by disyllabic equivalents.

Drawing on large historical as well as present-day corpora covering the 19th and 20th centuries (COHA, COCA, BNC; all accessed via the BYU interface), we portray some asymmetrical diachronic shifts and synchronic distributions of synonym pairs and trios across attributive and predicative uses. In doing so, a major issue will be the extraction of collocations of different strengths and the attempt to eliminate their influence on the distributions. We will discuss the genesis of collocations in view of their rhythmic and other properties as well as fluctuations in their frequencies.

A further set of analyses will shed more light on the nature of stress clash avoidance itself. The adjectives *shut* vs. *closed* are (nearly) synonymous and both monosyllabic. Yet, they differ substantially in their phonetic duration. This length difference can be hypothesized to account for the relative freedom with which *closed* occurs before initially stressed nouns, while the distribution of *shut* is significantly skewed: The stresses in *clósed éyes* are considerably further separated than those in *shút éyes*. The resultant syntactic asymmetry is as robust as in the case of *glad* vs. *happy*, suggesting that not only unstressed intervening syllables but any kind of temporal distance between stresses can satisfy the rhythmic requirement.

On a more tentative note, we will suggest that the influence of rhythm in general and on individual collocations in particular has waxed and waned in the diachronic perspective. For one thing, the loss of inflectional endings in adjectives has exacerbated rhythmic constellations in adjective-noun sequences and made them increasingly susceptible to rhythmic constraints. For another, a rhythmically motivated syntactic avoidance phenomenon may contribute to a secondary semantic specialization: Adjectives like *glad* and *shut*, which are now largely restricted to predicative uses, tend to take on temporal meanings while adjectives that alternate between attributive and predicative positions can be either temporal or characterizing (cf. Bolinger 1952, 1967).

References

Bolinger, D. L. 1952. Linear modification. *Papers of the Modern Language Association,* 67: 1117-44.

Bolinger, D. L. 1965. Pitch accent and sentence rhythm. In D. L. Bolinger. *Forms of English: Accent, Morpheme, Order*. Cambridge, Mass.: Harvard University Press, 139-80.

Bolinger, D. L. 1967. Adjectives in English: attribution and predication. *Lingua,* 18: 1-34.

Cruse, D. A. 2002. Paradigmatic relations of inclusion and identity III: synonymy. In D. A. Cruse, F. Hundsnurscher, M. Job & P. R. Lutzeier eds. *Lexikologie / Lexicology: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen / An International Handbook on the Nature and Structure of Words and Vocabularies*. Vol. I. Berlin: de Gruyter. 485-97.

Dolezal, F. 2013. Synonymy and sameness of meaning: an introductory note. *International Journal of Lexicography* 26(3): 255-59.

Ehret, K., C. Wolk & B. Szmrecsanyi. 2014. Quirky quadratures: on weight and rhythm as constraints in the genitive alternation. *English Language and Linguistics* 18(2): 263-303.

Schlüter, J. 2005. *Rhythmic Grammar. The Influence of Rhythm on Grammatical Variation and Change in English*. Berlin/New York: Mouton de Gruyter.

Shih, S., J. Grafmiller, R. Futrell & J. Bresnan. To appear. Rhythm's role in genitive construction choice in spoken English. In R. Vogel & R. van de Vijver eds. *Rhythm in Phonetics, Grammar and Cognition*. Berlin/New York: Mouton de Gruyter.

Low transitivity contexts as breeding grounds for novel verbs: An analysis of waxing and waning verbs

Ulrike Schneider & Britta Mondorf (University of Mainz)

The present paper investigates the role of transitivization and detransitivization in the waxing and waning of verbs. Defining transitivity with Hopper & Thompson (1980: 251) as the effectiveness with which the verbal action takes place, we set out to provide new findings to the question of whether low transitivity contexts provide a breeding ground for novel verbs.

Recent findings indicate that the death of – at least some – transitive verbs is accompanied by a detransitivization process (cf. Mondorf forthc.) which results in a low transitive stage before the verb finally dies out. Thus, formerly well-established causative verbs that used to be able to occur in highly transitive contexts eventually come to be restricted to low transitive contexts, i.e. uses in which the effect the verbal action can have on the object is reduced. Retreat to such low transitive contexts which score low on transitivity parameters such as affirmation (negation rather than affirmation), mode (irrealis rather than realis), participants (one participant rather than two), particles (particle rather than fully-fledged objects; cf. Cappelle 2007; Ogura 1994; Tenny 1994), etc. seems to be a common path for verbs leaving the language.

The present study is the first to test whether low transitivity contexts also serve as an entry-strategy or breeding ground for novel verbs. On the basis of diachronic corpora (e.g. Early English Prose Fiction, Eighteenth Century Fiction, Nineteenth Century Fiction, Corpus of Historical American English) we empirically investigate verbs that were introduced into the English language system after the 16th century, such as *derail*, *roam,* etc. measuring the development of their degree of transitivity on the basis of 16 parameters in order to provide a multifactorial analysis of their trajectory of change. The diachronic, corpus-based analyses of novel verbs provide new insights into the role of transitivity in language change.

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.

Hopper, P. J. & S. A. Thompson. 1980. Transitivity in grammar and discourse. *Language*, 56: 251-99.

Möhlig, R. & M. Klages. 2002. Detransitivization in the history of English from a semantic perspective. In T. Fanego, J. Pérez-Guerra & M. José López-Couso eds. *English Historical Syntax and Morphology*. Selected Papers from 11 ICEHL, Santiago de Compostela, 7-11 Sept 2000. Amsterdam: Benjamins, 231-54.

Mondorf, B. forthcoming. The detransitivization of causative *bring*. Special Issue on Support Strategies, *Language Variation and Change*.

Ogura, M. 1994. Grammatical choices in Old and Early Middle English: a choice between a simple verb, the prefix/particle-verb or verbpParticle combination, and the “auxiliary + infinitive” construction in Old and Early Middle English. In F. Fernández, M. Fuster & J. José Calvo eds. *English Historical Linguistics*. Papers from 7th International Conference on English Historical Linguistics. Valencia, 22-26 Sept 1992. Amsterdam: Benjamins, 119-29.

Tenny, C. 1994. *Aspectual Roles and the Syntax-Semantics Interface*. Dordrecht: Kluwer.

Words about animals

Alison Sealey (Lancaster University)

This paper presents preliminary results from a funded research project that has collected an extensive thematic corpus comprising a heterogeneous range of discourses which have in common a central concern with one or more (non-human) animals. The texts in the corpus include news stories, literature from organisations that campaign about animal-related issues (e.g. animal rights, pro-hunting groups, animal welfare), transcripts of wildlife documentaries, academic research articles, transcripts of interviews from the project with people whose work involves communicating about animals (e.g. broadcasters, scientists, meat producers), etc.

Among the motivations for the project is an interest in how corpus assisted discourse analysis can identify perceptions, values, attitudes and ideological orientations towards animals. This is worthy of investigation for its own sake, as it can help to illuminate the degree to which established ways of talking and writing are attuned to describing the rapidly changing environment in which humans and animals co-exist. In addition, it has wider implications, since it deals with the depiction of entities that are sentient but are not themselves directly involved in the production of discourse. In this way, the texts in this corpus contrast with those - often studied by (critical) discourse analysts - which represent human groups (e.g. women, immigrants), who may themselves be the authors of discourses in their own right.

This presentation will explain how the corpus was constructed, to include contemporary texts (produced between 1995 and 2015) in British English, as far as possible, and how parameters were devised to generate search terms so as to identify relevant texts for inclusion in the corpus.

It will go on to demonstrate some of the lexical patterns found in:

- the naming and classifying of different kinds of animal, with reference to both 'scientific' and 'folk' taxonomies
- the descriptions associated with various animals (e.g. by attributive and predicative adjectives)
- figurative expressions associated with a range of animals
- the processes represented, as animals act and are acted upon
- the roles of animals in these processes (e.g. as agent, goal)
- words representing the cognitive, emotional and social behaviour of animals.

The presentation will conclude by drawing out the relevance for corpus assisted lexical and discourse analysis of the findings of this study. It will consider implications both for the specific topic (discourse about animals) and for the broader ontological and epistemological issues raised by the project.

References

'People', 'Products', 'Pests', 'Pets': the discursive representation of animals. Leverhulme Trust Project Grant RPG-2013-063. https://animaldiscourse.wordpress.com.

Reconstructing language contact in the antebellum South: African-American Vernacular English and White Southern Vernacular English

Lucia Siebers (University of Regensburg)

While previous studies (Schneider 1989, Kautzsch 2002) on African American English (AAE) have yielded important insights into the evolution of this variety in the second half of the nineteenth and the beginning of the twentieth century, still very little is known about how English evolved among this community in the antebellum period. This paper sets out to report on new data from the Corpus of Older African American Letters, containing more than 1,500 semi-literate letters written by African Americans between 1763 and 1910. An earlier study on letters from this corpus provided evidence for the Northern Subject Rule and showed that there is relatively little regional variation in subject-verb concord in the 1880s and 1890s even in highly diverse sociohistorical settings (Siebers 2015). Thus the question emerges whether this was the result of continuing homogenization and levelling in the course of the nineteenth century (cf. Winford 1997:336) and if so, whether we find a similar degree of homogenization with regard to other features. Hence the main aim of this paper is to examine the amount of variability regarding subject-verb concord and past time reference with a focus on letters written in the first half of the nineteenth century.

While the focus of this article will be on AAE, the results will also be compared to the Southern Plantation Overseer Corpus (Schneider and Montgomery 2001, Trüb 2006), which contains similar semi-literate letters and - both in terms of region and period studied - nicely matches the AAE data. Although it is generally agreed that the development of AAE is inextricably linked to that of Southern American English, still very little is known about the relationship of earlier forms of both varieties (this holds particularly true for the first half of the nineteenth century). According to Bailey (2001:53), this is among other reasons due to the lack of emphasis on the sociohistorical context of AAE and the absence of suitable data for both varieties. With the availability of the two corpora, it is hoped that such a much needed comparison sheds further light on the amount of "linguistic variability in earlier black-white sociolinguistic relations" (Schneider 2007:268).

References

Bailey, G. 2001. The relationship between African American Vernacular English and White Vernaculars in the American South: A sociocultural history and some phonological evidence. In S. Lanehart ed. *Sociocultural and Historical Contexts of African American English*. Amsterdam: Benjamins, 53-92.

Kautzsch, A. 2002. *The Historical Evolution of Early African American English. A Comparison of Early Sources*. Berlin, New York: Mouton de Gruyter.

Schneider, E. W. 1989. *American Earlier Black English: Morphological and Syntactic Variables*. Tuscaloosa: Alabama University Press.

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

Schneider, E. W. & M. Montgomery 2001. On the trail of early nonstandard grammar: an electronic corpus of Southern U.S. Antbellum overseers' letters. *American Speech*, 76: 388-410.

Siebers, L. 2015. Assessing heterogeneity. In A. Auer, D. Schreier & R. J. Watts eds. *Letter Writing and Language Change*. Cambridge: Cambridge University Press, 240-63.

Trüb, R. 2006. Nonstandard verbal paradigms in Earlier White Southern American English. *American Speech*, 81: 250-65.
Winford, D. 1997. On the origins of African American Vernacular English. A creolist perspective. Part 1: Sociohistorical background. *Diachronica* XIV: 305-44.

PP complements in AdvP structure

Urszula Skrzypik (University of Chester)

Huddleston and Pullum argue that prepositions, unlike most adverbs, license complements, which constitutes one of the authors' generalisations that supports their boundary between the two word classes (2002: 58, 571). They say that the adverbs that license complements are 'virtually limited' to the set in (1) (the prepositions licensed by the adverb heads are in brackets):

(1) *separately (from); independently (of, from); similarly (to); equally (with); differently (than, from, to); analogously (to); comparably (to); identically (to); concomitantly (with); concurrently (with); consistently (with); simultaneously (with)* (Huddleston and Pullum 2002: 571)

Given that only a handful of *·ly* adverbs license complements, Huddleston and Pullum assert that if a word not marked with the *·ly* suffix licenses a complement, it is not an adverb but a preposition (or belongs to some other class) (2002: 571). Hence (for example) *near* is said to be a preposition as it licenses a *to* complement (ibid.: 609).

The aim of this paper is to critically discuss the authors' criterion and to provide corpus and web data that challenge the generalisation in question. Thus I argue that considering that the ability to license complements pertains to the prototypical (i.e. *·ly*) adverbs, it is not convincing to expel, on these grounds, from the adverb category the items other than *·ly* members (e.g. *close*, *near*) that can also license complements. This is because there is a general tendency that defining features of a particular word class spread from the core to the periphery (Aarts 2007: 97−111). If the core adverbs license complements, it is to be expected that a number of non-central members would behave in a similar way. Moreover, Huddleston and Pullum's generalisation is inconsistent with their argumentation in other parts of *The Cambridge Grammar*. The authors argue that '[d]ifference in complementation doesn't justify a primary part-of-speech distinction'; yet, 'it is the pattern of complementation that provides the most general criterion for distinguishing prepositions from adverbs' (Huddleston and Pullum 2002: 604, 1012).

This paper also shows that Huddleston and Pullum (2002) underestimate the ability of *·ly* adverbs to license PP complements. I checked complementation patterns of a large number of adjectives in LDOCE (2003), and subsequently investigated whether the adjectival complementation patterns extend to the corresponding AdvPs by searching the British National Corpus and the web for relevant examples. These methods enabled me to expand the set in (1) significantly. As a case in point, consider (2).

(2) Under the MacSharry proposals Scotland would suffer [disproportionately to not only other EC countries but other countries within the United Kingdom]. (BNC: HHX: 4130)

In (2) the AdvP headed by *disproportionately* takes the PP *to not only other EC countries but other countries within the United Kingdom* as its complement as evident from the difficulty of replacing the preposition *to* with other prepositions. Therefore, the PP *headed* by *to* qualifies as complement of *disproportionately* by the licensing criterion. Further (corpus- or web-derived) adverbs that license PP complements are (for example) *remotely (from), adjacently (to), vertically (to), symbolically (of), ir-/relevantly (to), representatively (of), sympathetically (to), synonymously (with), responsively (to), proportionately (to), in-/compatibly (with), inconsistently (with), inseparably (from), irrespectively (of).*

In conclusion, given that Huddleston and Pullum's argumentation is not always plausible and that there exist a considerable number of prototypical adverbs that license PP complements, the authors' assertion that a word not ending in the *·ly* suffix though licensing a complement is not an adverb appears to be doubtful. It is reasonable to consider words such as *near* (as in *near to the station*) as adverbs, thereby questioning Huddleston and Pullum's boundary between adverbs and prepositions.

References

Aarts, B. 2007. *Syntactic Gradience: The Nature of Grammatical Indeterminacy*. Oxford: Oxford University Press.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Hoffmann, S., S. Evert, N. Smith, D. Lee & Y. Berglund Prytz, 2008. *Corpus Linguistics with BNCweb: a Practical Guide*. Frankfurt am Main: Peter Lang.

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

Longman Dictionary of Contemporary English (LDOCE), 4th edn. 2003. D. Summers (director). Harlow: Pearson Education.

The emergence of temporal subordinators across inner- and outer-circle varieties of English

Adam Smith (Macquarie University)

Contemporary grammars such as Biber et al. (1999) acknowledge the class of complex subordinators introducing adverbial clauses such as *as far as, the way (that), on condition (that).* One adverbial category that offers a particularly rich range of such constructions is that expressing time. Edgren (1971) notes a set of temporal "phrasal conjunctions" including *the time, the day, the minute, the moment* (sometimes preceded by a preposition, or followed by the relative pronoun *that/when*). As noted by Peters (2012), when there is either a preceding prep and/or a following relative pronoun, a phrase like *(at) the instant* remains an adverbial constituent of the first clause, with *that* as relativizer for the second clause, as in:

(1) It all changes **(at) the instant that** he realizes…

Their grammatical role as subordinators is clearcut in their elliptical form, without the preceding preposition or following relative pronoun:

(2) The ringer will stop **the instant** the phone is picked up

A study of a set of temporal phrasal adverbials in the International Corpus of English (ICE) (Smith, 2014), found evidence for three of them in particular as complex subordinators: *the instant*, *the minute*, *the moment*. Based on evidence from the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), these constructions were found to be more common in British English (BrE) than in American English (AmE). This paper will present an analysis of these three possibly emerging complex subordinators, using data from the co-contemporary and equivalent-sized BrE and AmE sections of the Corpus of Global Web-based English (GloWbE) to test the hypothesis that they are more established in BrE than in AmE. It has been noted (Schneider, 2007) that lexicogrammatical innovations are characteristic of postcolonial Englishes, and therefore the ICE and GloWbE corpora of two outer-circle varieties – Hong Kong and Philippines English – have been interrogated. The relative current state of emergence of these temporal subordinators can thus be compared across two inner-circle varieties and two outer-circle varieties that have developed from them (Hong Kong from BrE and Philippines from AmE).

Frequencies for the elliptical forms have been compared with the non-elliptical forms, to establish the relative state of emergence of the forms with a likely subordinative role, as opposed to those that may be functioning as a prepositional phrase. In addition, data showing the frequency of clause-initial occurrences of each item will demonstrate how often they occur in contexts where their function as a subordinator can be shown to be unambiguous (Smith, 2014). While generic comparisons cannot be attempted due to the different composition of the corpora used, this study will demonstrate the effectiveness of the GloWbE corpora in providing the evidence on different stages of development of complex subordinators found in the ICE corpora. In addition, it will provide further evidence for the grammaticalisation of a set of temporal subordinators, in both inner- and outer-circle varieties, that have previously been considered no more than marginal.

References

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

Edgren, E. 1971. *Temporal Clauses in English*. Acta Universitatis Upsaliensis (Studia Anglistica Usaliensa 9). Uppsala: Uppsala University.

Peters, P. 2012. Emergent conjunctions. In S. Chevalier & T. Honegger eds. *Words, Words, Words: Philology and Beyond: Festschrift for Andreas Fischer on the Occasion of his 65th birthday.* 127-44.

Schneider, E. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press

Smith, A. 2014. Newly emerging coordinators in spoken and written English. *Australian Journal of* Linguistics, 24 (1): 118-38.

Charting ongoing change: The emergent complex subordinators *(at/from) the moment (that)* and *for/out of/in fear (that)/of*

Adam Smith (Macquarie University)
Lot Brems (Université de Liège)
Kristin Davidse (KU Leuven)

There is no agreement yet in the literature on the recognition criteria for emergent complex subordinators (henceforth CSs), even though it is generally accepted that a mix of syntactic, semantic and categorial change is involved. To shed light on this still largely intractable mix, we will focus on CSs which derive from preposition + noun + postmodifying clause, viz. *at/from the moment that* and *for/out of/in fear that/of*. They exemplify two types of constructions that diachronic studies have shown to be potential sources of CSs, noun + relative clause, e.g. *the while that*, Hopper & Traugott (2003), and noun + complement clause, e.g. *in order that/to* Łęcki & Nykiel (forthc.). By studying the emergent CS uses of the proposed strings in a range of subordinate constructions, we aim to formulate broad generalizations about the crucial factors in the gradient change from source to target.

We will develop a multifactorial analytical framework to categorize contextualized data in relation to the subordinator reading as: not allowing it, allowing it as one possible, but not exclusive, reading, or allowing a subordinator reading only, i.e. as Diewald's (2006) lexical, critical or isolating contexts.

Syntactically, the grammaticalization process involves 'upranking' (Halliday 1994) of the secondary clause from embedded postmodifier of the noun to a hypotactic clause linked to its matrix by a complex subordinator incorporating that noun (Brems & Davidse 2010, Smith 2014, Nykiel 2013). Isolating contexts of CS *the moment* where identified by Smith (2014) as occurring in sentence-initial position, e.g. (1), and when punctuation suggests a separate tone unit, e.g. (2), as pointed out by Peters (2012: 136) for *the way*.

(1) *The moment* I heard your snooty twang on the phone I knew it was love. (ICE-AUS)
(2) Here Paul tells … about how the late Derek Bell … changed his life, *the moment* he fell instantly in love with his wife (WB)

Both context types instantiate what Verstraete (2007) calls 'free' subordination, where both the subordinated clause and the matrix have an information structure with a focus of their own. We will systematically investigate these contexts and their information distribution (also when not signalled by a comma) as enabling CS uses of *the moment* and *for fear that/of* (3)-(4). Contexts structurally ambiguous between adjunct and subordinated clause (Huddleston & Pullum 2002: 1012-4) occur when the preposition + noun + clause string is informationally integrated with the matrix as its information focus, as in (5)-(6).

(3) *for fear that* enemy bombers would use the transmitters for guidance, the service was closed down. (ICE-GB)
(4) I let him take risks, *for fear* of smothering him. (WB)

(5) Peking had proclaimed that we prisoners would be executed *the moment* our army advanced (WB)

(6) I kept staring at it, resisting the urge to blink *for fear* it would vanish (WB)

We will complement this syntax- and pragmatics-based classification of ambiguous and non-ambiguous contexts with a semantic analysis, distinguishing the reference to a specific moment or fear construed as an abstract entity in the lexical uses from the relational meanings of the subordinator uses, 'as soon as' / 'when' and 'lest' respectively. This will in turn be correlated with the formal indices of *decategorialization* (Hopper 1991), i.e. deletion of elements associated with the NP uses, such as [- preposition], [- determiner] [- relative or complementizer].

The proposed qualitative and quantitative analysis will be applied to exhaustive extractions on the strings in question from ICE-GB, ICE-AUS and ICE-USA, building on the findings of Smith (2014) for these data, supplemented with register-balanced extractions from the UK, US and Oz subcorpora of Wordbanks*Online*. We will assess whether variety or mode appear to promote the grammaticalization process.

References

Brems, L. & K. Davidse. 2010. Complex subordinators derived from noun complement clauses: grammaticalization and paradigmaticity. *Acta Linguistica Hafniensia*, 42: 101-16.

Diewald, G. 2006. Context types in grammaticalization as constructions. *Constructions*. http://elanguage.net/journals/index.php/constructions/article/viewArticle/24

Halliday, M.A.K. 1985. *An Introduction to Functional Grammar*. 2nd ed. London: Arnold.

Hopper, P. 1991. On some principles of grammaticalization. In E.Traugott & B. Heine eds. *Approaches to Grammaticalization*. Vol 1. Amsterdam: Benjamins, 17-36.

Hopper, P. & E. Traugott. 2003. *Grammaticalization*. 2nd ed. Cambridge: CUP.

Huddleston, R. & G. Pullum. 2002. *The Cambridge Grammar of the English Language.* Cambridge: CUP.

Nykiel, J. 2014. Grammaticalization reconciled: functionalist and minimalist insights into the development of purpose subordinators in English. *Language Sciences*, 42: 1-14.

Łęcki, A. & J. Nykiel. forthcoming. Grammaticalisation of the English prepositional conjunction *in order to/that*. In H. Cuyckens, L. Ghesquière & D. Van Olmen eds. *Aspects of Grammaticalization: (Inter)Subjectification, Analogy and Unidirectionality*. Berlin: Mouton de Gruyter.

Peters, P. 2012 Emergent conjunctions. In S. Chevalier & T. Honegger eds. *Words, Words, Words, Philology and Beyond: Festschrift for Andreas Fischer on the Occasion of his 65th birthday.* Tübingen: Francke Verlag. 127-44.

Smith, A. 2014. Newly emerging subordinators in spoken/written English. *Australian Journal of Linguistics*, 34: 118-38.

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

Verstraete, J. C. 2007. *Rethinking the Coordinate-Subordinate Dichotomy. Interpersonal Grammar and the Analysis of Adverbial Clauses in English*. Berlin: Mouton.

Was and *Were* in the English counterfactuals: Semantic distinction and usage variation

Myounghyoun Song (Seoul National University)

Corpus linguists (Johansson (1988), Peters (1998), Turner (1980)) and general grammarians (Quirk, et. al. (1985), Huddleston & Pullum (2002)) investigated the usage variation between *was* and *were* in the English counterfactual conditionals (henceforth, *was counterfactuals* and *were counterfactuals*), assuming that there is no semantic difference between the two forms. Even semanticists (Lewis (1973), Iatridou (2000), Arregui (2008)) have focused their theory about counterfactuals on the truth conditions, paying little attention to the semantic distinction between the two counterfactuals, which is shown in (1).

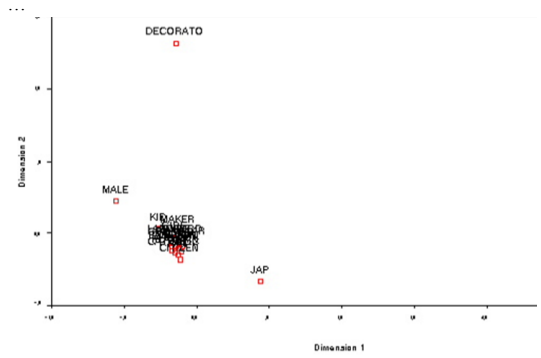
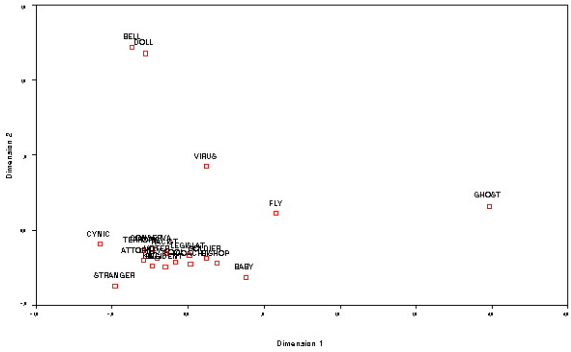
(1) a. If I *was/were a bell, I'd go ding, dong (COCA_Spoken_2010).
b. If I was/were president, this is what I would do (COCA_Spoken_1995).

We see in (1) that the *were counterfactual* is only acceptable in (a) when the speaker (or the human being) receives the counterfactual property of being a 'bell' and the *was counterfactual* is favored in (b) if the speaker is a president candidate while the *were counterfactual* is favored if he is a comedian. The aim of this paper is first to investigate the semantic distinction between the two counterfactuals and then to go over the variation in usage between the two counterfactuals when they show semantic overlaps.

In this paper, the semantic difference between the two counterfactuals is investigated in two ways that correspond to the examples in (1). First, with a list of the predicate nouns following the past copular forms separated into three categories with respect to their contextual distribution, as in (2),

predicate nouns occurring only with *was*	predicate nouns occurring with both *was* and *were*	predicate nouns occurring only with *were*
burglar, Catholic, citizen, criminal, decorator, doctor, fan, girl, intern, jap, kid, lady, andlord, layman, maker, male, patient, player, wife, …	candidate, cop, director, journalist, killer, police, politician, student, teenager, manager, governor, lunatic, mother, musician, lawyer, woman, man,…	attorney, baby, bell, bishop, coach, conservative, cynic, doll, family, fly, fool, ghost, king, legislator, racist, resident, soldier, stranger, terrorist, virus, voter,…

the semantic similarity between the discriminating collocates (Gries, 2001) in (a) and (c) respectively is measured by the Resnik similarity metric (Resnik, 1995) and graphically represented with a multidimensional scaling plot (Young & Hamer, 2001), as in the figures below, which reveals the fact that predicate nouns after *was* are semantically closer to each other (mostly, hyponyms of person) than those after *were* (almost everything including physical and abstract things).

Semantic dissimilarity between predicate nouns after 'was'	Semantic dissimilarity between predicate nouns after 'were'
	

Secondly, the contextual overlaps (Rubenstein & Goodenough, 1965) in (2b) are differentiated by conceptual differences between the subject in *was counterfactuals* and that in *were counterfactuals* (e.g. I as a president and I as a comedian in (1b)), which will reveal the fact that *was counterfactuals* are favored to portray the situation as 'possible to be actualized in the future' while were counterfactuals are favored to portray the situation as impossible to be actualized in the future.

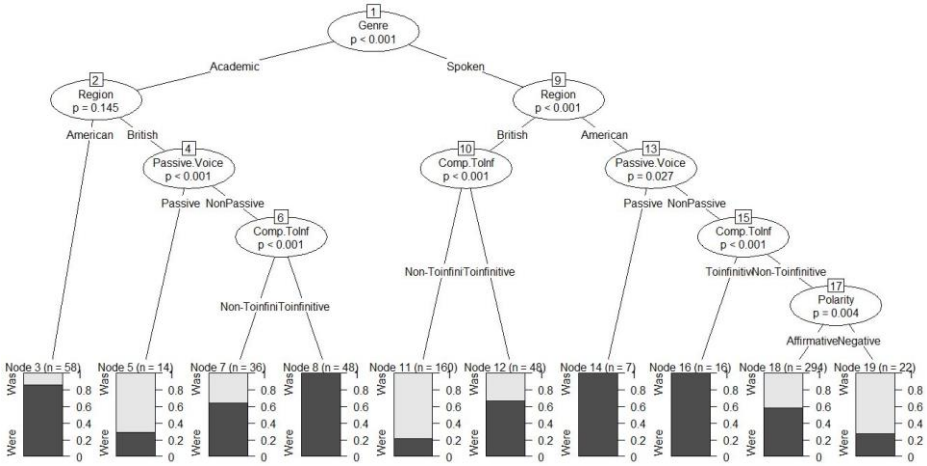
While the two forms show semantic difference with regard to actualizability in the future, they seem to show variation in usage without any semantic difference when they refer to the vicarious identity or property in the present time or they both function as auxiliary verbs taking appropriate verbal forms, as in (3).

(3) a. If I were/was <u>you</u>, I wouldn't be out walking in this weather. (personal pronoun)
b. It would be better if I were/was <u>dead</u>! (adjective)
c. I would be thinking along those same lines if I were/was <u>in</u> your shoes. (preposition)
d. I would have a very difficult time if I were/was <u>running</u> for reelection. (present participle)
e. If I were/was <u>forced</u> to sell quickly, I'd take a big loss. (passive participle)
f. If I were/was <u>to come</u> right out with it in two sentences, you wouldn't believe me. (to infinitive)

Also the two counterfactuals co-vary with such other linguistic environments as 'polarity', 'inversion' and 'subject persons', as well as such extra-linguistic contexts as regions (British vs. American) and genres (written vs. spoken). In this paper, the usage patterns underlying the variation between the two forms are uncovered with an enhanced dataset from the relatively recent corpora (COCA and BYU-BNC), as seen in the table below.

Group	Factors	*were* counterfactuals		*was* counterfactuals		Total	
		No.	%	No.	%	No.	%
Region	American	251	63.2	146	36.8	397	56.5
	British	141	46.1	165	53.9	306	43.5
Genre	spoken	267	48.8	280	51.2	547	77.8
	written	125	80.1	31	19.9	156	22.2
Subject Person	1st	216	54.5	180	45.5	396	56.3
	3rd	176	57.3	131	42.7	307	43.7
Polarity	affirmative	372	56.9	282	43.1	654	93.0
	negative	20	40.8	29	59.2	49	7.0
Complementation	NP(pronoun)	116	60.7	75	39.3	191	27.2
	AP	44	43.1	58	56.9	102	14.5
	PP	28	36.4	49	63.6	77	11.0
	present participle	68	48.2	73	51.8	141	20.1
	passive participle	40	50.0	40	50.0	80	11.4
	to Infinitive	96	85.7	16	14.3	112	15.9
TOTAL		**392**	**55.8**	**311**	**44.2**	**703**	**100**

The data were analyzed with GoldVarb X to identify statistically significant factors to the variation (Taglimonte, 2006) and the *party* package to R software to provide a model that can predict the usage of the two counterfactuals in various contexts, as shown below:



The implications of the findings put forth so far are twofold. First, *was counterfactuals* indicate that the antecedent situation is actualizable in the future while *were counterfactuals* suggest that the antecedent situation is not. Second, the variation patterns in usage suggest that *were counterfactuals* pervade the academic American English while in British Academic English they are more favored with *to-infinitive* complements but less favored in the passive constructions and the progressive constructions. *Was counterfactuals*, on the other hand, are generally preferred by the British speakers while *were counterfactuals* are preferred by the American speakers, except in the negative constructions.

References

Arregui, A. 2008. On the role of past tense in resolving similarity in counterfactuals. A. Grønn ed. *Proceedings of SuB12*, Oslo: 17-31.

Gries, 2001. A corpus-linguistic analysis of English *–ic* vs. *–ical* adjectives, *ICAME Journal*, 25: 65-108.
Huddleston, R. & Pullum, G. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Iatridou, S. 2000. The grammatical ingredients of counterfactuality. *Linguistic Inquiry* 31: 231-70.
Johansson, S. 1988. The subjunctive in British English and American English. *ICAME Journal* 12: 27-36.
Lewis, D. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.
Peters, P. 1998. The survival of the subjunctive. *English World-Wide*, 19, 1: 87-103.
Quirk, R., Greenbaum, S., Leech, G., & Startvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
Resnik, P. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.
Rubenstein, H & Goodenough, J.B. 1965. Contextual correlates of synonymy. Vo1 8 (10): 627-33.
Tagliamonte, S.A. 2006. *Analyzing Sociolinguistic Variation*. New York: Cambridge University Press.
Turner, J.F. 1980. The marked subjunctive in contemporary English. *Studia Neophilologica*, 52:2, 271-77.
Turney, P. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001).*
Young. F. W. & R. M. Hamer. 1994. *Theory and Applications of Multidimensional Scaling.* Hillsdale, NJ: Eribaum Associates.

Let's try AND/TO figure this out! Using spoken vernacular corpora to inform explanation

Sali A. Tagliamonte & Marisa Brook (University of Toronto)

The verb TRY can introduce a complement clause either with TO or AND (Lind 1983, Quirk et al. 1985:507, Biber et al. 1999:738). The construction with AND predates the verb's modern meaning, to 'attempt' or 'endeavour', which arose via a semantic shift from earlier meanings of TRY ('examine' or 'test the strength/value of') (Tottie 2012). Recent analyses argue that there is no current semantic difference between TRY TO and TRY AND (Lind 1983, Gries and Stefanowitsch 2004, Ross 2013, Tottie 2012). Indeed, both variants can occur in the same speaker and same stretch of discourse, as in (1), (see also Lind 1983:559), suggesting that TRY TO versus TRY AND is a linguistic variable (Labov 1972).

(1) People have to TRY TO figure out what you're doing and the person that guesses it, the person that's drawing it has to run and tag- TRY AND tag them before they reach back to their spot. (Triana Selowsky, F 24, Temiskaming Shores, Ontario)

A review of the extant literature presents a mixed picture. On one hand, there is a British/American contrast for spoken data: TRY AND 71 percent in the UK versus 24 percent in the US (Hommerberg and Tottie 2007:48). On the other hand, there is a strong register difference in British English: TRY AND is most frequent in spoken data. Yet TRY TO is the dominant written form, dating back to the 17th century (Tottie 2012:209-210). What explains this development? According to Tottie (2012), when the semantic reanalysis

occurred, the infinitive marker TO became a possible alternative. This option took off in written data: Google N-gram results show TRY TO progressively gaining ground ever since (Ross 2013). Lexical effects may underlie this development since BE is reported to resist TRY AND (Lind 1983:562).

In this paper we probe the TRY TO/TRY AND alternation using sociolinguistically stratified corpora of vernacular speech from Canadian and British communities, thus enabling us to corroborate cross-variety differences, establish whether there is change in progress and test for how speakers deploy the two variants in usage.

Quantitative analysis of over 1400 tokens of TRY AND/TO confirms robust variation between TRY AND and TRY TO in British and Canadian English, as well as a distinct regional difference consistent with Hommerberg and Tottie (2007). TRY AND is both the incoming variant in British English and the most frequent, accounting for 72 percent of the data. In Canadian English, the variable is stable in apparent time, and TRY AND appears only 30 percent of the time. In neither variety is the variation constrained by social factors. Rather, the lexical effect dominates. As predicted by Lind (1983), BE is the most resistant to TRY AND, and this is consistent across both corpora.

We suggest that ongoing variation between TRY AND and TRY TO is the synchronic reflex of reanalysis. Because the original meaning of TRY AND ('test') was not compatible with BE, there has been ongoing resistance across communities to the construction TRY AND BE over and above the effect of community or register. Thus, an important finding of this study is the extent to which particular collocations contribute to the overall patterning, persisting across centuries. Further, this research corroborates accumulating research advocating the key role of the lexicon in grammatical variation (Torres Cacoullos and Walker 2009; Poplack 1992).

References

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

Carden, G. & D. Pesetsky. 1977. Double-verb constructions, markedness, and a fake coordination. *Papers from the 13th Annual Meeting of the Chicago Linguistic Society*. Chicago: University of Chicago, 82-92.

Faarlund, J. T. & P. Trudgill. 1999. Pseudo-coordination in English: the 'try and' problem. *Zeitschrift für Anglistik und Amerikanistik*, 47(2): 210-13.

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

Hommerberg, C. & G. Tottie. 2007. Try to or try and? *ICAME Journal*, 31: 43-62.

Labov, W. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lind, Å. 1983. The variant forms *try and/try to*. *English Studies*, 64(6): 550-63.

Poplack, S. 1992. The inherent variability of the French subjunctive. In C. Lauefer & T. A Morgan eds. *Theoretical Studies in Romance Linguistics*. Amsterdam: Benjamins, 235-63.

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

Ross, D. 2013. Dialectal variation and diachronic development of *try*-complementation. *Studies in the Linguistic Sciences: Illinois Working Papers*, 38(2). 108-47.

Torres-Cacoullos, R. & J. A. Walker. 2009. The present of the English future: grammatical variation and collocations in discourse. *Language*, 85(2): 321-54.
Tottie, G. 2012. On the history of *try* with verbal complements. In S. Chevalier & T. Honegger eds. *Words, Words, Words: Philology and Beyond: Festschrift for Andreas Fischer on the Occasion of his 65th birthday*, Tübingen: Narr Francke Attempto, 199-214.

The complexity of pronoun omission: An analysis based on ICE Singapore and ICE India

Iván Tamaredo, J. Carlos Acuña-Fariña & Teresa Fanego (University of Santiago de Compostela)

One of the central assumptions of twentieth-century linguistics was that all languages are equal with respect to their overall grammatical complexity. In recent years, however, several studies have shown that languages differ in complexity (see, for example, Miestamo et al. 2008), and metrics have been designed to measure complexity in an empirical way. One such metric is that of Hawkins (2004), which measures the cost of processing individual constructions for language users. Hawkins argues that processing linguistic forms and their grammatical features requires effort, and therefore it is easier for speakers to use structures that allow them to convey the same message using fewer forms.

The focus of the present paper is pronoun omission, a grammatical feature that makes structures easier to process according to Hawkins' metric, because it reduces the formal complexity of the structure. However, this is so only if the referent of the omitted pronoun can be recovered from discourse by using formal and/or non-formal cues. In a previous study (Tamaredo 2014), it was found that in Indian and Singapore English, two varieties that show a high frequency of pronoun omission (cf. Kortmann & Lunkenheimer 2013), a considerable percentage of all cases of pronoun omission occurred when referents were highly "accessible". The aim of the present study is to see whether, in addition to accessibility (on this notion, see especially Ariel 1994, 2001), formal cues can also facilitate the omission of pronouns in these two varieties. The assumption is that formal marking on verbs, in the form of agreement suffixes attached to the verbs, provides information about their arguments, and that this may make the search for the appropriate referents of omitted pronouns easier, thus reducing processing costs. Huddleston and Pullum et al. (2002: 94) distinguish between three main types of verbs in English: lexical verbs, non-modal auxiliaries (*be, have, do*), and modal auxiliaries. On the assumption that more formal marking on verbs facilitates pronoun omission, omitted pronouns should occur more frequently in subject position with lexical verbs and non-modal auxiliaries than with modal auxiliaries, since the former two both have paradigms in which more forms of the verb agree in person and number with the subject. In English, most verbs only agree with third person singular subjects and only in the present tense, but the absence of agreement is also informative because it signals that the subject is not third person singular. In object position, however, the frequency of omitted pronouns should be similar for the three types of verb, because English does not have object-verb agreement.

In order to test this hypothesis, instances of omitted and overt pronouns were retrieved from the Indian and Singaporean components of the *International Corpus of English*, identifying the type of verb (lexical, non-modal auxiliary or modal auxiliary) with which they occurred. The results for omitted pronouns were then compared with those for overt ones, as a means of determining the frequency with which pronouns are dropped with each type of verb.

Findings from a pilot study show that there are more instances of omitted pronouns in subject position with lexical verbs than with either non-modal or modal auxiliaries. In addition, the distribution of omitted versus overt pronouns in object position is the same for the three types of verb. These results are in partial agreement with the hypothesis that more formal marking on verbs increases the ratio of pronoun omission. However, this is not the case with non-modal auxiliaries. Our preliminary findings thus suggest that formal marking on the verb may be a factor which influences pronoun omission in Indian English and Singapore English, although it appears to be a secondary one. This opens up the possibility that formal marking interacts with frequency effects (cf. Bybee & Hopper 2001), including the process whereby a string of words that co-occur regularly (e.g. pronoun + non-modal auxiliary) becomes stored in memory as a complex unit.

References

Ariel, M. 1994. Interpreting anaphoric expressions: a cognitive versus a pragmatic approach. *Journal of Linguistics*, 30: 3-42.

Ariel, M. 2001. Accessibility theory: an overview. In T. Sanders, J. Schliperoord & W. Spooren eds. Text Representation, Amsterdam: Benjamins, 29-87.

Bybee, J. & P. Hopper eds. 2001. *Frequency and the Emergence of Linguistic Structure.* Amsterdam: Benjamins.

Hawkins, J. A. 2004. *Efficiency and Complexity in Grammars.* Oxford: Oxford University Press.

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

Kortmann, B. & K. Lunkenheimer eds. 2013. *The Electronic World Atlas of Varieties of English*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://www.ewave-atlas.org/ Accessed 6 Dec 2014.

Miestamo, M., K. Sinnemäki & F. Karlsson eds. 2008. *Language Complexity: Typology, Contact, Change*. Amsterdam: Benjamins.

Tamaredo, I. 2014. Pronoun drop in contact varieties of English. Paper delivered at the *38th Conference of AEDEAN*, University of Alcalá de Henares.

Data sources

ICE-IND = *International Corpus of English - the Indian Component*. 2002. Project coordinated by Prof. S. V. Shastri at Shivaji University, India, and Prof. Dr. Gerhard Leitner at Freie Universität Berlin, Germany. http://ice-corpora.net/ice/download.htm.

ICE-SIN = *International Corpus of English - the Singaporean Component*. 2002. Project coordinated by Prof. Paroo Nihilani, Dr Ni Yibin, Dr Anne Pakir and Dr Vincent Ooi at The National University of Singapore, Singapore. http://ice-corpora.net/ice/download.htm.

A corpus-based production-oriented legal English Dictionary for non-native English speaking law students: To enable competition on an equal footing

Shinichiro Torikai & Masayuki Tamaruya (Rikkyo University)

English legal discourse is notorious for its incomprehensibility. It is especially so for non-native speakers of English. The non-native English speaking students who aim to pass the bar exam or seek a law degree in an English speaking country must spend a tremendous amount of time and energy to overcome this difficulty before being able to compete with English speaking students.

General English dictionaries are helpful for understanding the general use of English for non-technical communication, but they seldom list legal technical terms or provide illustration from actual usage in legal discourse. Legal dictionaries such as Black's Law Dictionary do list legal technical terms exhaustively and with precision, but the entries rarely include authentic examples of legal discourse.

With the support of a Japanese government research grant, the presenters compiled a corpus-based production-oriented legal English dictionary for non-native English speaking students. The corpus was created using 1,242,656 words from UK Supreme Court Judgments, 1,139,952 words from U.S. Supreme Court Judgments, 1,135,346 words from five UK law journals, and 1,141,299 words from 14 U.S. law journals. In addition, the Oxford English Dictionary (1989) was used to identify 2,806 words with the keyword *law* and their 3,183 corresponding definitions. From this exercise, we found that the difficulties of legal English, at least for non-native English speakers, can be grouped into three categories.

Category 1 represents general English words that are polysemous, having a legal meaning in addition to a more common definition. Such words as *act, action, case,* and *hold*, are all general English words used for everyday conversation, but once they are used in legal discourse, their meanings are completely different, which causes confusion for non-native English speaking students.

Category 2 represents legal synonyms. Most lay people do not know or are unable to explain the difference among *homicide, murder* and *manslaughter*. Moreover, how many non-native English speaking students can use *decision, decree, finding, judgment, ruling,* sentence and *verdict* correctly? General English dictionaries and legal dictionaries do not explain well how these legal technical terms are used in legal discourse.

Category 3 represents basic English words that are used as a pivot to express complicated legal processes. We found that legal expressions are often structurally fixed with regard to sentence structure, and certain patterns are repeatedly used. In such cases basic verbs are used as a pivot to combine key components of the legal case. This can be illustrated by using the word file.

(1) Merrill Lynch instead **filed** an interpleader **action under** 28 U.S.C. § 1335. (US JDG)

(2) Respondent **filed** his own **suit under** the Jones Act…, **alleging**… (US JDG)

(3) A taxpayer must **file** an administrative **claim with** the Internal Revenue Service before **filing suit against** the Government… (US JDG)

All the above structural patterns using *file* as a pivot follow the same basic structure (see Table 1):

Table 1: Example Structural Pattern Based Upon the Pivot Word "File"

[a party]	pivot	[legal procedure]	[nature of procedure]	[the substance of argument]
respondent	**file**	a motion	in (court)	seeking…
defendant		a petition	with (agency/court)	requesting…
petitioner		suit	against (defendant)	alleging…
attorney		a lawsuit	under (law/regulation)	challenging…
		a complaint	on the ground that…	asserting…
		a brief	on ground	claiming…
		an action	pursuant to…	contending…
		a claim		to challenge…

After a corpus-based analysis we concluded that these three categories of general English words, legal synonyms, and basic English words used as pivots are the keys for non-native English students to comprehend and produce legal discourse. We will demonstrate how they are effective by showing some sample descriptions of our corpus-based legal English dictionary and examples from our legal corpora.

References

Garner, B. A. *et al*. 2004. *Black's Law Dictionary*. St. Paul, Minnesota: West, a Thompson business.

Gartner, B. 2011. *Garner's Dictionary of Legal Usage.* Oxford: Oxford University Press.

Gilmour L. *et al*. eds. 1995. *Collins English Thesaurus.The Ultimate Wordfinder from A to Z*. Glasgow: Harper Collins.

Gove, P. *et al*. eds. 1984. *Webster's New Dictionary of Synonyms*. Springfield, Massachusetts: Merriam-Webster.

Hori, M. 2009. *Introduction to Collocation Studies in English*. Tokyo: Kenkyusha.

Ichikawa, S. *et al*. eds. 1995. *The Kenkyusha Dictionary of English Collocations*. Tokyo: Kenkyusha.

Inoue, N. & Akano, I. eds. 2013. *The Wisdom English-Japanese Dictionary*. Tokyo: Sanseido.

Kipher, B. A., *et al*. eds. 2001. *Roget's International Thesaurus.* New York: Harper Collins.

Konishi, T. & Minamide, K. *et al*. eds. 2006. *Genius English-Japanese Dictionary*. Tokyo: Taishukan.

Mayer, M. *et al*. eds. 2009. *Longman Dictionary of Contemporary English.* Essex: Person Education Ltd.

Simpson, J. A. & E. S. C. Weiner eds. 1989. *Oxford English Dictionary*. Oxford: Oxford University Press.

Sinclair, J. *et al*. eds. 2009. *Collins Cobuild Advanced Learner's English Dictionary*. Glasgow: Harper Collins.

Takebayashi, S. *et al*. eds. 2002. *Kenkyusha's New English-Japanese Dictionary*. Tokyo: Kenkyusha.

Tanaka, H. *et al.* eds. 1991. *Dictionary of Anglo-American Law.* Tokyo: Tokyo University Press.

Urdang, L. ed. 1997. *The Oxford Thesaurus.* Oxford: Oxford University Press.

Urdang, L. *et al*. eds. 1986. *Longman Synonym Dictionary*. Essex: Longman Group.

Waite, M. *et al*. eds. 2002. *Concise Oxford Thesaurus.* Oxford: Oxford University Press.

Waite, M. *et al*. eds. 2009. *Oxford Thesaurus of English.* Oxford: Oxford University Press.

From pause to word? *Uh* and *um* in written language

Gunnel Tottie (The University of Zürich)

The vocalizations usually transliterated as *uh* and *um* in American English and *er* and *erm* in British English have been been variously classified as dysfluencies, hesitations, hesitation markers, fillers, filled pauses, etc.

Only a few linguists have classified them as words, most notably the psycholinguists Clark and Fox Tree (2002), in a study based on the London-Lund Corpus. They argue that UHM are not *symptoms* of hesitation but words, which are *signals* of upcoming delays (i.e. silent pauses), This claim has been refuted by O'Connell and Kowal (2005), based on interviews with Hillary Clinton, and Tottie (2015), based on the Santa Barbara Corpus of Spoken American English (SBCSAE). These authors found that most instances of *uh* and *um* are not followed by delays, and also that most silent pauses are not announced, and that therefore the announcement of upcoming delays cannot be used as support for the status of *uh* and *um* as words.

However, Clark and Fox Tree point out in a footnote that *uh* and *um* are used deliberately in chatrooms and emails, and that this supports their classification as words. There is even stronger support for regarding *uh* and *um* as words when they appear in print – in headlines and commentaries in newspapers and magazines – as in (1) – (6) from the Zurich newspaper corpus, BNC and COCA.

(1) **Obama is more, um, seasoned**. Barack Obama's …closely shorn hair appears to be increasingly gray. (*Washington Post* 2008)

(2) **An ode to opera's, uh, operation**. As…Baroque-era composers become increasingly popular, more people wonder about the castrati – the emasculated singers … (*L. A. Times* 2005)

(3) (Context: Ben Affleck goes to sleep during a performance of Shakespeare's *As You Like It*) The Oscar-winning screenwriter and actor seemed to be, **um** – how can we put this delicately? – meditating during most of the first act. (*Boston Globe 1999*)

(4) …she is currently looking for an accountant who is willing to do her books in return for, **er,** payment in kind. Her number's in the phonebook. (*Accountant*, BNC)

(5) …orthopedists who treat aging hikers don't have an ironic, **um,** bone in their bodies. (*Backpacker*, COCA)

(6) …Senator Richard Shelby of Alabama… claimed that …"The market will view these firms as … implicitly backed by the government." **Um, senator**, the market already views those firms as having implicit government backing, because they do... (Paul Krugman, Op-Ed Column in *The New York Times*)

In (1) – (5) the function of *uh* and *um* is that of ironic euphemism, "putting it delicately," and in (6), *um* is used for disagreement in a vocative. *Uh* and *um* can thus function as words and are used as stance markers in writing, at least in "agile" genres (Hundt & Mair). Early results indicate that this usage has progressed further in American English than in British English,

and that there are register differences within and between the two varieties. In order to establish whether *uh* and *um* can be regarded as words in speech as well, I also examine spoken corpora (BNC and the SBCSAE) for similar uses; early results indicate that *uh* and *um* operate on a cline between (filled) pause and word in spoken language. Moreover, some functions of *uh* and *um* seem to only appear in writing, which parallels the development of *well* in writing discussed by Rühlemann and Hilpert (MS).

References

Clark, H. H. &. J. E. Fox Tree, 2002. Using *uh* and *um* in spontaneous speaking. *Cognition,* 84: 73–111.

Hundt, M. & C. Mair. 1999. 'Agile' and 'uptight' genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics,* 4, 221–42.

O'Connell, D. C. & S. Kowal. 2005. Uh and Um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research,* 34: 555–76.

Rühlemann, C. & M. Hilpert. MS. Colloquialization in journalistic writing: the case of inserts with a focus on *well*.

Tottie, G. 2015. *Uh* and *um* in British and American English: Are they words? Evidence from co-occurrence with pauses. In N. Dion, A. Lapierre & R. Torres Cacoullos eds. *Linguistic Variation: Confronting Fact and Theory.* New York: Routledge, 38–54.

Word order of reporting and reported clauses in language-contact situations: A comparison of translated English and ESL writing

Bertus van Rooy (North-West University)
Haidee Kruger (Macquarie University/North-West University)

A central assumption of construction grammar is that grammatical constructions emerge out of repeated use; in other words, frequency of use is central to the emergence of constructional schemas (Bybee 2010). Language-contact situations have the potential to confront the language user with competing frequencies for a similar or cognate grammatical construction in two or more languages. One such example is the word order of the reporting and reported clause in English and Afrikaans, as illustrated by example (1), where the reporting clause precedes the reported clause, and (2), where the reported clause precedes the reporting clause, from Biber *et al*. (1999:922):

(1) She said: "Elderly people often have smaller groups of friends and family to support them."
(2) "We may all be famous, then," said he.

In English fiction, according to Biber *et al*. (1999), the final position is used in about two-thirds of all cases of direct speech, but in news reportage only slightly more than half of the instances of direct speech have the reporting clause in final position. In Afrikaans, data from the *Taalkommissiekorpus* [Corpus of the Language Commission] reveal a consistent preference of about 70% for the final position in both news and fiction. When considering indirect speech in Afrikaans, the preference shifts overwhelmingly to the initial position.

Researchers in translation studies (e.g. De Sutter & Van de Velde, 2008) have examined a number of other constructions where frequency distributions and/or contexts of use differ across languages, and accounted for the observations in terms of constraints on translated language, such as normativity and the priming effects of the source text or even the source language more generally. Relatively independently, researchers studying language acquisition, such as Van Vuuren (2013), have investigated similar constructional pairs across two languages, and explained residual non-nativeness in the use of advanced learners in terms of the transfer of the frequency of use of constructions from the first language to the second language. Lantsyák and Heltai (2012) propose that translated language and bilingual communication may share a number of linguistic properties due to the shared constraint of bilingual processing operating in both. However, Malkiel (2009:228) cautions that as far as cognate words in two languages are concerned, bilinguals happily make use of the "free vocabulary", whereas translators often avoid the use of cognate items, presumably for fear of using false friends.

The frequency, and especially differences in frequency, of grammatical constructions plays an important part in trying to resolve some of the theoretical issues in the two language-contact situations: writing in a second language and translation. This paper examines reporting clauses in English and Afrikaans comparatively in native control corpora of the two languages, as well as corpora of English produced by native speakers of Afrikaans and translations from Afrikaans to English in order to advance our understanding of what happens in language-contact situations in the settings of second-language English writing and translation into English. At the same time, we attempt to advance our understanding of the role of frequency in the emergence of grammatical constructions, using a corpus-based analysis of language-contact data.

References

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

Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

De Sutter, G. & M. Van de Velde. 2008. Do the mechanisms that govern syntactic choices differ between original and translated language? A corpus-based translation study of PP extraposition in Dutch and German. *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2008)*. http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/de_Sutter_and_de_Velde.pdf (Accessed 3 September 2014.)

Lanstyák, I. & P. Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures*, 13(1): 99-121.

Malkiel, B.. 2009. Translation as a decision process: evidence from cognates. *Babel*, 55(3): 228-43.

Van Vuuren, S. 2013. Information structural transfer in advanced Dutch EFL writing: a cross-linguistic longitudinal study. *Linguistics in the Netherlands 2013*: 173-87.

"Talk talk, not just small talk": Exploring contrastive focus reduplication with the help of corpora

Bianca Widlitzki (University of Giessen)

This paper discusses contrastive focus reduplication (CR) in contemporary English. CR is a type of full reduplication whose first element bears contrastive stress. Broadly speaking, its function is to constrain the meaning of an item by "narrowing down the range of appropriate referents" in potentially ambiguous contexts (Ghomeshi et al. 2004: 317). *Talk talk* in (1) is an example:

(1) My Dad and I actually talked. Like, not just small talk, but talk talk. It was very nice indeed. (Blog Authorship Corpus 825029)

While several studies have discussed CR based on arbitrarily selected examples (Ghomeshi et al. 2004, Song and Lee 2011), there is almost no corpus-based research. Hohenhaus (2004), using a corpus of fictional TV and movie dialogues, is one exception.

The present study extends corpus-based research on CR to a different genre and medium by using the Blog Authorship Corpus (Schler et al. 2006), which contains blog entries from the year 2004. Existing scholarship is especially concerned with the semantics of CR. While this aspect is addressed in the present study, the form of CR (e.g. What bases are frequently used?) and its linguistic context (e.g. What is contrasted with CR?) are also analyzed. Finally, the Blog Authorship Corpus allows studying sociolinguistic factors as it includes information on bloggers' gender and age.

The corpus has two further advantages. The relatively informal nature of blogs and the enormous size of the corpus (c. 140 million words) increase the chances of retrieving this colloquial and infrequent phenomenon. In the end, roughly 40 million words, a little more than a quarter of the entire corpus, were searched with specialist software that identified any reduplicated string of letters (Reduplication Finder; Fessl 2006), ultimately yielding 113 instances of CR.

Analysis of the data adds valuable details to our understanding of CR, both in terms of its structural and semantic properties as well as concerning its users. For instance, it lends support to Hohenhaus's (2004: 311-312) claim that most bases for CR are morphologically simple nouns or adjectives. However, adverbs turn out to be more frequent than previously thought: they constitute c. 17% of the CR tokens in the blogs, whereas Hohenhaus (2004: 310) did not discover any in his corpus. Although the rarity of CR makes statistical statements on the effects of age and gender difficult, the results nevertheless indicate an overuse among female writers and younger bloggers.

To gather more examples for the analysis of the linguistic context of CR, the Corpus of Contemporary American English and the Corpus of American Soaps (Davies 2008-, 2012) were consulted. As their web interfaces do not support searches for repeated strings, they were searched for known CRs, which produced 200 additional tokens. It emerges that the

unreduplicated base – whether on its own or as part of a compound (as *small talk* in 1) – is not necessarily required as a contrast to CR if there is other contextual information (cf. 2):

(2) I live in a mobile home park. I'm not poor poor, but I am in the lower class tax bracket. (Blog Authorship Corpus 1936076)

In conclusion, CR appears to be a very flexible phenomenon that also occurs in contexts other than spoken dialogue, namely in blogs, and, though rare, is productive in contemporary English.

References

Davies, M. 2008-. *The Corpus of Contemporary American English: 450 million words, 1990-present*. http://corpus.byu.edu/coca. Accessed 1 Dec 2014.

Davies, M. 2012. *The Corpus of American Soap Operas: 100 million words, 2001-2012*. http://corpus2.byu.edu/soap/. Accessed 1 Dec 2014.

Fessl, A. 2006. Technical Documentation of Reduplication Finder 1.1. Graz: Department of Linguistics, University of Graz. http://reduplication.uni-graz.at/.

Ghomeshi, J., R. Jackendoff, N. Rosen & K. Russell. 2004. Contrastive focus reduplication in English (The Salad-Salad Paper). *Natural Language & Linguistic Theory* 22 (2): 307–357.

Hohenhaus, P. 2004. Identical constituent compounding: a corpus-based study. *Folia Linguistica*, 38(3-4): 297-32.

Schler, J., M. Koppel, S. Argamon & J. Pennebaker. 2006. Effects of age and gender on blogging. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*: 199-205. http://lingcog.iit.edu/doc/springsymp-blogs-final.pdf. Accessed 28 Nov 2014.

Song, M. H. & C. Lee. 2011. CF-reduplication: dynamic prototypes and contrastive focus effects. In N. Ashton, A. Chereches & D. Lutz eds. *Proceedings of the 21st Semantics and Linguistic Theory Conference (SALT 21)*, 444-462. http://elanguage.net/journals/salt/issue/view/324. Accessed 27 Nov 2014.

Paraphrases and reporting phrases in learner summaries: The development of intertextual competence in intermediate L2 writing

Leonie Wiemeyer (University of Bremen, Germany)
Barbara Ann Güldenring (Philipps-University Marburg, Germany)

The aim of this paper is to explore the development of intertextual competence via a quasi-longitudinal corpus-based study of paraphrases in the summaries of German intermediate learners of English. For this purpose, high-school student summaries from the *Marburg Corpus of Intermediate Learner English* (MILE) (Kreyer 2015) are analysed in order to trace differences in the use of paraphrases and accompanying reporting phrases in three consecutive school years.

Intertextual competence, i.e. the ability to appropriately reference discipline-specific discourse, is a central aspect of academic writing and thus is often fostered as a core component of students' writing abilities in various educational contexts. Thus, appropriate intertextuality, i.e. critical assessment and use of source texts, often forms the basis for measuring a learner's competence in academic writing (Shaw & Pecorari 2013: A1). Using

appropriate academic language is especially challenging in a foreign language, particularly when learners have to refer to other texts in their writing (Abasi & Akbari 2008, Thompson, Morton & Storch 2013; Davis 2013), as not every type of intertextuality is evaluated positively (see, for example, Crocker & Shaw 2002 for a discussion of the evaluation of patchwriting by educators and linguists).

Experienced academic writers use paraphrases in order to not only keep the textual similarities between their texts and the source material to a minimum, but also to encode their own assessment of the reproduced statements (Hyland 2002; Uccelli, Dobbs & Scott (2013). In school, paraphrasing is often taught as one strategy to prevent plagiarism. When paraphrasing, learners are required to reproduce other authors' texts in their own words. Keck (2006, 2014) defines attempted paraphrases (APs) as passages reproducing the source text that contain at least one lexical alteration of the source text, e.g. a synonym (see 1). These structures, which are often introduced by optional reporting phrases (underlined in 2), are an important intertextual feature of written summaries.

(1) Original: Working so many hours does not leave much time for hobbies and sports.

Student paraphrase: They work a lot of time, there isn't any time for hobbies. (MILE, 000500122)

(2) <u>First the author explains</u> which generation has the bigest problems about being illegal immigrants. (MILE, 005800053)

As shown by Keck (2006), the paraphrases of L2 writers at university level differ significantly from those by their native-speaker peers as the former draw much more heavily on the vocabulary of the paraphrased text.

In order to explore the development of intertextual competence of learners of English at high school level, APs and reporting phrases were identified in 52 summaries from grades 9, 11, and 12 from MILE. The goal was to determine whether a decreasing reliance on source text vocabulary and structures can be observed. For this purpose, learner paraphrases were categorised using Keck's (2006) taxonomy of paraphrase types based on similarities between source and student texts. The taxonomy comprises four types, namely Near Copies, Minimal Revisions, Moderate Revisions, and Substantial Revisions, which are delineated from each other by the percentage of words identical with the source text excerpt ("unique links") and the presence of additional features such as synonym substitution, nominalisation, and clause structure revision. In order to investigate the intermediate learners' growing intertextual competence, additional analyses were conducted focussing on the restructuring of source text information and the semantics and complexity of reporting phrases, for which purpose reporting phrases were categorised based on their syntactic structure and coded for their evaluative function.

The study shows growing intertextual competence from grade 9 to 12 as more advanced learners use less source text vocabulary and rely more on their own expressions. They also tend to use reporting phrases as rhetorical devices to improve their summaries. The reporting

phrases of grade 12 pupils are much more syntactically complex and semantically varied than those of the younger cohorts. However, significant inter-learner variability exists in all three cohorts with regard to paraphrase and reporting phrase structure. While some learners actively restructure the text and use their own words, others rely heavily on exact copies and often fail to correctly reproduce the information in the source text, possibly due to a lack of reading comprehension.

References

Abasi, A. R. & N. Akbari. 2008. Are we encouraging patchwriting? Reconsidering the role of the pedagogical context in ESL student writers' transgressive intertextuality. *English for Specific Purposes*, 27: 267-84.

Crocker, J. & P. Shaw. 2002. Research student and supervisor evaluation of intertextuality practices. *Hermes Journal of Linguistics*, 28: 39-58.

Davis, M. 2013. The development of source use by international postgraduate students. *Journal of English for Academic Purposes*, 12(2): 125-35.

Hyland, K. 2002. Activity and evaluation. Reporting practices in academic writing. In J. Flowerdew ed. *Academic Discourse*. Harlow: Longman, 115-30.

Keck, C. 2006. The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing* 15(4): 261-78.

Keck, C. 2014. Copying, paraphrasing, and academic writing development: A re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing*, 25: 4-22.

Kreyer, R. 2015. The Marburg Corpus of Intermediate Learner English (MILE). In M. Callies & S. Götz eds. *Learner Corpora in Language Testing and Assessment*. Amsterdam: Benjamins.

Shaw, P. & D. Pecorari. 2013. Source use in academic writing: An introduction to the special issue. *Journal of English for Academic Purposes* 12 (2): A1–A3.

Thompson, C., J. Morton & N. Storch. 2013. Where from, who, why and how? A study of the use of sources by first year L2 university students. *Journal of English for Academic Purposes* 12(2): 99-109.

Uccelli, P., C. Dobbs, L. Christina & J. Scott. 2013. Mastering academic language. Organization and stance in the persuasive writing of high school students. *Written Communication,* 30(1): 36-62.

Software demonstrations

Large-scale time-sensitive semantic analysis of historical corpora

Paul Rayson, Alistair Baron, Scott Piao & Steve Wattam (UCREL research centre, School of Computing and Communications, Lancaster University, UK)

Previous attempts to apply automatic semantic analysis to Early Modern English (EmodE) corpora have employed existing taxonomies developed for modern corpora such as the USAS tagset (Rayson et al, 2004). However, this fails to account for significant meaning and vocabulary shifts over time. What is required is a broad coverage taxonomy combined with historically sensitive meaning categories. The Historical Thesaurus of English (http://historicalthesaurus.arts.gla.ac.uk) developed at the University of Glasgow over forty years is such a system. It provides a high-quality semantic lexical database containing around 800,000 entries manually classified into 236,000 thesaurus categories arranged in a hierarchical structure. A key challenge is to scale up the semantic disambiguation in USAS, currently based on a smaller semantic field taxonomy of 232 tags for modern English, to that of the Historical Thesaurus with much finer grained distinctions. A smaller set of four thousand thematic codes devised at Glasgow and arranged at an intermediate level in the hierarchy can also be applied in order to produce semantically tagged output. In this software demonstration, we will show the new Historical Thesaurus Semantic Tagger (HTST) which uses the full set of categories, thematic codes and USAS tags. The user can also enter the date of a text and the software employs dating information in the thesaurus to help it choose more appropriate categories. In addition, given the links from the Historical Thesaurus to the entries in the Oxford English Dictionary (OED), we are able to draw on further information from the senses, definitions and example sentences in the OED in order to assist in the ranking of contextually appropriate thesaurus and thematic codes.

A second significant challenge in the application of corpus and computational linguistics methods to EmodE corpora is historical spelling variation which has been shown to significantly affect their accuracy and robustness (Archer et al, 2003; Rayson et al, 2007; Baron et al, 2009). Following the development of the Variant Detector (VARD) software (Baron and Rayson, 2008), this problem can be addressed by inserting modern equivalents alongside historical variants which can then be tagged, counted and searched for with appropriate software. We are undertaking a large crowdsourcing exercise which will permit the large-scale manual training of time-sensitive models for matching historical spelling variants. These models can then be applied to our corpora to achieve more accurate results. Moreover, we can make use of variant spelling and dating information in the OED to improve the accuracy and coverage of VARD.

The final challenge in this enterprise is the large-scale and heterogeneous nature of the corpora and metadata. In particular, we have indexed the transcribed portion of the Early English Books Online (EEBO-TCP; http://www.textcreationpartnership.org/tcp-eebo/), over one billion words. The latest release from the Text Creation Partnership (TCP) in November 2014 brought the transcribed portion to 53,830 books. For corpus software, running such a collection through a pipeline consisting of historical spelling variant normalisation (VARD),

part-of-speech tagging (CLAWS), semantic tagging (USAS and HTST) followed by indexing (in tools such as CQPweb and Wmatrix) requires significant computational resources.

The software demonstration will show how we have been able to overcome all three of these challenges, demonstrate how accurate such processes are and signpost directions for future work. The HTST is available to use at: http://phlox.lancs.ac.uk/ucrel/semtagger/english.

Acknowledgements

This work took place in the Semantic Annotation and Mark-Up for Enhancing Lexical Searches (SAMUELS) project (http://www.gla.ac.uk/samuels/) funded by the Arts and Humanities Research Council in conjunction with the Economic and Social Research Council (grant reference AH/L010062/1), January 2014 to March 2015. The VARD crowdsourcing experiment is funded by JISC in the UK.

References

Archer, D., McEnery, T., Rayson, P., Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In *Proceedings of the Corpus Linguistics 2003 conference.* UCREL technical paper number 16. UCREL, Lancaster University, pp. 22-31.

Baron, A. and Rayson, P. (2008). VARD2: A tool for dealing with spelling varia9on in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, 22nd May 2008.

Baron, A., Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. In *Anglistik: International Journal of English Studies*, 20 (1), pp. 41-67.

Rayson, P., Archer, D., Piao, S., & McEnery, A. M. (2004). The UCREL semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop*, Lisbon, Portugal, 2004. (pp. 7-12). Lisbon.

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics 2007*, July 27-30, University of Birmingham, UK.

How to get more out of a corpus: A generic multilayer corpus query interface

Simon Sauer (Humboldt-Universität zu Berlin)
John M. Kirk (Technische Universität Dresden)

Many corpora are made available only in the format(s) used for compiling and annotating. While some formats require corpus users to download and install specific software, others are very generic and can be used with any text editor. Naturally, however, these formats are optimized for the corpus builders' own research interests and might make investigating other research questions very difficult even when the data as such is applicable. Metadata are often not incorporated into the corpus proper and are only available within the corpus's documentation.

Web-based corpus interfaces can alleviate most of these issues. Developing and maintaining a separate interface for each corpus, however, is time-consuming and expensive. Generic

interfaces such as CQPweb (1), on the other hand, are often restricted to simple token-based annotations such as part-of-speech tags.

ANNIS (2) is an open source, cross platform, web browser-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation. It is format- and theory-agnostic as all data are modelled as abstract nodes and edges. ANNIS is compatible with a multitude of formats through the SaltNPepper meta model and conversion framework (3). Queries are formulated in a powerful query language, which allows complex queries across different annotation types and even across corpora. Unicode and regular expressions are fully supported. Query hits can be displayed in a variety of independent visualization modules, such as a classic key word in context view, a grid-style view that supports both token-based annotations and annotation spans, a document view where annotations are represented by colour-coded highlighting, arches for dependency relations, trees for syntactic structures and others. Aligned audio and video data can be played back by clicking on any token or annotation.

The underlying multilayer standoff architecture allows for more than one 'basic' text layer and for annotations completely independent from each other. For example, the diachronic corpus RIDGES (4) features a manuscript-near transcription, a further transcription using only contemporary characters and resolving graphical issues such as word-internal line breaks as well as a normalized transcription in today's orthography. Graphical annotations such as paragraphs, pages, headings, or margins are based on the manuscript-near transcription whereas part-of-speech tags refer to the normalized transcription. Metadata can be queried just like any other annotation, so you can easily form ad-hoc subcorpora by limiting your results to for example manuscripts from before 1630.

Besides ANNIS's features in general, this paper will demonstrate its benefits on the concrete example of ICE- and SPICE-Ireland (5), which hitherto were only available in a linear text-based format, with metadata being restricted to the corpus handbook. In ANNIS, the spoken subcorpus displays speakers on separate layers, so overlaps are graphically represented and much easier to recognize. All annotations are also on separate layers, which significantly improves both legibility of the actual text and the ability to query specific annotations.

References

(1) CQPweb: http://cqpweb.lancs.ac.uk
A. Hardie. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*. 17 (3): 380-409. http://www.lancaster.ac.uk/staff/hardiea/cqpweb-paper.pdf

(2) ANNIS: http://annis-tools.org
A. Zeldes, J. Ritz, A. Lüdeling & Ch. Chiarcos. 2009. ANNIS: A search tool for multi-layer annotated corpora. In M. Mahlberg, V. González-Díaz & C. Smith eds. *Proceedings of Corpus Linguistics 2009*. http://edoc.hu-berlin.de/docviews/abstract.php?id=36996

(3) SaltNPepper: http://korpling.german.hu-berlin.de/saltnpepper
F. Zipser & L. Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In G. Budin, L. Romary, T. Declerck & P. Wittenburg eds. *LREC 2010 Workshop, Proceedings, W4: Language Resource and Language Technology Standards*. Paris: ELRA. http://hal.inria.fr/inria-00527799

(4) RIDGES: http://korpling.german.hu-berlin.de/ridges/index_en.html
(5) ICE- and SPICE-Ireland: http://ice-corpora.net/ice/iceire.htm

Multilingualiser 1.0

Jukka Tyrkkö (University of Tampere)

In this talk I will discuss and demonstrate the identification and tagging of non-English words and phrases using the new tool *Multilingualiser* 1.0. The tool has been designed to meet the needs of the *Multilingual Practices in the History of Written English* project at the University of Tampere, but it can be of use for any project with the need to identify, tag and retrieve specific sets of words or code-switches in corpora regardless of the languages involved. *Multilingualiser* is not a language-identification tool but rather a tool for identifying individual words and short passages in texts predominantly written in a single language.

The tool includes three separate functions: a query tool, a tagger and a manual tag editor. The query tool makes use of dictionary look-up, collocate analysis, character n-grams and date evaluation. Dictionaries for French, Latin, Italian and German are included, but user-defined dictionaries can also be used if needed. Collocate analysis or the analysis of words within a user-defined window of successive target words helps identify words not included in the dictionaries by analysing contexts where a single such words appears in a string of foreign words of an identified foreign language. The same method also helps disambiguate between closely related languages in cases where dictionary look-up gives two possible source languages such as Italian and Latin. Character n-grams allow the identification of words that contain character strings that are highly infrequent in English; this is particularly useful in the case of code-switching to non-Germanic and non-Italic languages. An in-progress date-evaluation feature makes use of OED first datings harvested from the database of the *Historical Thesaurus of the Oxford English Dictionary*, courtesy of the HTE team in Glasgow, flagging items such as nominalisations which may be common today but were not in frequent use at the time the source text was published.

The tagger tool appends either underscore or XML tags to individual words, retaining headers and pre-existing tagging if desired. In XML mode, the tagging follows TEI P5 guidelines using BCP 47 language codes. Additional functions include adding sequential ID values to lexical items and transforming underscore-format tags such as traditional corpus linguistic POS-tags into XML attributes. The resulting new file can be exported or analysed in-tool for chunks or sequences of foreign items, which can be exported as a CSV file with a user-defined amount of window span. The tag editor allows the user to refine the query results, see a visual representation of the dispersion of code-switches and to make universal corrections to the tagging if necessary.

Multilingualiser 1.0 will be available free of charge for OS X, PC and Linux.

Pragmatic annotation & analysis in DART

Martin Weisser (Guangdong University of Foreign Studies)

Until recently, conducting medium or large-scale research in pragmatics has been an extremely time-consuming effort, due to the inherent difficulties in identifying and annotating speech acts at a reasonably fine-grained and yet sufficiently generic level.

With the recent release of the Dialogue Annotation And Research Tool (DART), such an enterprise has become considerably easier, making it possible to create pragmatically annotated corpora of orthographically transcribed dialogues by automatically identifying some 80+ speech-act combinations, along with information about the syntactic category, semantics (topic), semantico-pragmatic features (IFIDs or modes), as well as surface polarity, of each c-unit.

This extended software demonstration will introduce the basic ideas behind the design of the tool, its simple, yet extensive, annotation scheme, automatic annotation capabilities, and variety of (pre- & post-)editing and analysis options, such as the built-in concordancer, n-gram analysis, etc., using illustrative material from a number of different corpora.

The software itself, along with an extensive manual, can be obtained free of charge under GPL 3 licence from the presenter's website at http://martinweisser.org.

Work-in-progress reports

A preliminary approach to 'ephemeral' conditional adverbial subordinators in the history of English

Cristina Blanco-García (University of Santiago de Compostela)

In the course of the Early Modern English period (sixteenth and seventeenth centuries), conditional subordination increased considerably and the use of conditional subordinators became more differentiated and regulated (cf., among others, Görlach 1991: 122; Barber 1997: 205; Rissanen 1999: 189). Early Modern English also represents a key period in the development of what Kortmann (1997: 301) has denominated 'ephemeral' adverbial subordinators, i.e. those that were added to the inventory of adverbial connectives in Late Middle English or, more commonly, Early Modern English and which lasted shortly or became obsolete or highly restricted beyond these periods. As Kortmann notes (1997: 333), conditional subordinators constitute no exception among adverbial connectives, since their number expanded greatly in Early Modern English and many of them dropped out of use after this period.

Over the last twenty years or so, there has been a substantial body of research on adverbial subordination and adverbial subordinators from various approaches (cf., among others, the monographs by Kortmann (1997), Pérez Quintero (2002) and Lenker (2010) and the collective volume edited by Lenker and Meurman-Solin (2007)). However, a comprehensive approach to so-called 'ephemeral' adverbial subordinators *per se* has not been attempted yet. In this context, the aim of this work-in-progress report is to offer a preliminary approach to 'ephemeral' connectives within the domain of conditional subordination, among others: *if so be (that, as), be it so, if case be (that),* and *say (that).* Examples (1) and (2) illustrate the use of two of these 'ephemeral' conditional markers:

(1) **If so be** the Lord will be with me, then I shall bee able to driue them out **.** (1611 Bible (King James) Josh. xiv 12; OED s.v. if, conj. and n. 8.e)

(2) **If case be that** yow wyll that I schall send them ouyr to yow or to ony oder for yow, send me worde and it schal be don.It ought nat to be applied. (1482 J. Dalton Let. 27 Jan. In Cely Let. (1975) 129 R.; OED s.v. case n.1. P6.a)

Given the low frequency of these subordinators, in order to provide a comprehensive view of the development of this set of 'ephemeral' conditional connectives, data are drawn from various corpora. The *Penn-Helsinki Parsed Corpus of Early Modern English* and the *Penn-Helsinki Parsed Corpus of Modern British English* will be used as a base line, complemented with data from the *Oxford English Dictionary*, the *Middle English Dictionary*, and their quotation databases. Since conditional clauses seem to be particularly common in scientific texts in the Early Modern English period (Claridge 2007: 233), data from the aforementioned corpora will be supplemented with evidence from the *Corpus of Early English Medical Writing*. WordSmith Tools will be used to extract all the potentially relevant examples, in all their possible variant forms. These will then be manually disambiguated in order to discard all those cases in which the forms under investigation are not used as adverbial subordinators.

The analysis will pay attention to the following issues: (i) the overall quantitative distribution of the subordinators in the material across time; (ii) the process of grammaticalisation (cf. Hopper and Traugott 2003) undergone by the connectives under study; (iii) the verbal mood favoured by these subordinators in the sub-clause; (iv) the position of conditional adverbial clauses introduced by these connectives in relation to the main clause; (v) the combination of these subordinators with the 'pleonastic' *that* (cf. Beal 1988; Rissanen 1999); and (vi) their use in different text types and by different authors.

Sources and tools

Kroch, A., B. Santorini, & A. Diertani. 2004. *Penn-Helsinki Parsed Corpus of Early Modern English*. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html.

Kroch, A., B. Santorini & A. Diertani. 2010. *Penn Parsed Corpus of Modern British English.* http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html.

Middle English Dictionary. http://ets.umdl.umich.edu/m/med.

Oxford English Dictionary Online http://www.oed.com.

Scott, M. 2014. *WordSmith Tools version 6*, Liverpool: Lexical Analysis Software.

Taavitsainen, I., P. Pahta, & M. Mäkinen (compilers). 2010. *Corpus of Early English Medical Writing*. CD-ROM. Amsterdam: John Benjamins.

References

Beal, J. 1988. Goodbye to All 'That'? The history and present behaviour of optional 'that'. In G. Nixon & J Honey eds. *An Historic Tongue. Studies in English Linguistics in Memory of Barbara Strang.* London & New York: Routledge, 49-66.

Barber, C. 1997. *Early Modern English.* 2nd *ed*. Edinburgh: Edinburgh University Press.

Claridge, C. 2007. Conditionals in Early Modern English texts. In U. Lenker & A. Meurman-Solin eds. *Connectives in the History of English*. Amsterdam & Philadelphia: John Benjamins, 229-54.

Görlach, M. 1991. *Introduction to Early Modern English.* Cambridge: Cambridge University Press.

Hopper, P. J. & E. C. Traugott. 2003. *Grammaticalization*. 2nd ed. Cambridge: Cambridge University Press.

Kortmann, B. 1997. *Adverbial Subordination: A Typology and History of Adverbial Subordinators Based on European Languages*. Berlin & New York: Mouton de Gruyter.

Lenker, U. 2010. *Argument and Rhetoric. Adverbial Connectors in the History of English*. Berlin: Mouton de Gruyter.

Lenker, U & A. Meurman-Solin eds. 2007. *Connectives in the History of English*. Amsterdam & Philadelphia: John Benjamins.

Pérez Quintero, M. J. 2002. *Adverbial Subordination in English. A Functional Approach*. Amsterdam & New York: Rodopi.

Rissanen, M. 1999. Syntax. In R.M. Hogg ed. *The Cambridge History of the English Language*. Vol. 3. Cambridge: Cambridge University Press: 187-331.

Introducing the Corpus of Business English Correspondence (COBEC): A resource for the lexicon and pragmatics of Business English

Rachele De Felice (UCL) & Emma Moreton (Coventry University)

This paper provides an overview of COBEC – a new corpus of business correspondence in English, currently under development. COBEC contains approximately 50,000 emails and

several thousand letters dating from 2000 to 2006. The corpus data was donated by a telecommunications company that has now ceased operating. At present, we are developing the email component of the corpus. All emails are written in English. Approximately 30% of the contributors are non-native speakers of English, making COBEC a useful resource for exploring issues of intercultural communication. The corpus contains internal correspondence between the organisation's various operations in Europe and the US, as well as external correspondence between the organisation, its customers and suppliers. Information about the sender and the recipient – including role, sex and first language – is captured in the header alongside other contextual information such as whether the email is part of a chain, or contains attachments, and the relationship between correspondents, for instance.

This paper will describe the data making up the corpus, as well as some of the intended applications of the corpus. Challenges encountered in trying to capture the contextual information described above, in a TEI P5 compliant format, will be discussed, as well as the use of various natural language processing tools to semi-automate the annotation process, which will encompass contextual, linguistic, and discoursal features. In particular, we focus on the results of semi-automated pragmatic annotation of the data, using a tagging tool designed to identify the main speech act categories of request, commitment, expressive, and question.

We discuss the viability of this approach, and describe how it assists us in highlighting the variation present in the phraseology of speech acts in written business English. We present a case study centred on potentially face-threatening lexical items across several categories, including verbs (e.g. *want, need, must*), adverbs (e.g. *urgently, immediately, today*), and phrases (e.g. *as soon as possible, right away)*. Their presence in requests would not normally be expected, given the widely acknowledged tendency to mitigate rather than intensify this speech act, and their use would be considered potentially impolite for the hearer. Conversely, using these terms in commitments can achieve a positive effect, by making the speaker appear efficient and attentive to the hearer's needs; however, this carries the risk of threat to one's own face, as it can make the speaker's time and activities appear less important than the hearer's. Given the inherent risks in both types of speech acts, this paper addresses the following questions: (1) What contexts license the use of these 'expressions of urgency'? Is their use restricted to particular actions only? (2) Are they counterbalanced by other politeness markers such as mitigators?

Awareness of these usage patterns can benefit those learning how to use English in the workplace by highlighting expected norms of communication in a British English work environment. This study showcases the importance of resources such as COBEC, particularly when enriched with tagging and metadata information.

The ADJ + *enough* + resultative *that*-clause construction: Diachronic development and conditions of use

Signe Oksefjell Ebeling (University of Oslo)

This work-in-progress report takes two superficially similar structures as its starting point: the resultative (1) and the explanative (2) uses of the ADJ + *enough* + *that*-clause construction. Based on preliminary observations, my initial hypothesis is that the resultative type is on the increase, and my aim is to test this against diachronic corpus data.

(1) … we felt that the matter w was **important enough** that all members of this council should have an opportunity to debate it. (Mindt 2011: 139)

(2) It was **bad enough** that she'd fallen in love with the cold, glacial man she already knew him to be… (BNC/JY5 2744)

While the *that*-clause in a resultative construction, as in (1), reports a result, or consequence in Francis et al.'s (1998: 358; 362) terms, the *that*-clause in an explanative construction, as in (2), "provides an explanation in relation to the information given in the matrix clause" (Mindt 2011: 127). A formal difference between the two constructions is that the adverb *enough* is optional in the explanative construction and obligatory in the resultative construction.

In her study of adjectives complemented by *that*-clauses, Mindt (2011) found six instances of the resultative type in the British National Corpus (BNC). A search for the sequence "_AJ0 enough that" in BNC*Web* (cqp edition) returns 72 hits. It is assumed that the remaining 66 instances not explicitly discussed by Mindt are of the explanative type.

Being concerned with the potential increase in use of resultative ADJ *enough that*, the current study turns to the diachronic Corpus of Historical American English (COHA), spanning 1810-2009. A search for "[jj] enough that" in COHA seems to confirm the hypothesis that the sequence is on the rise, showing a relatively stable frequency of around 1-1.5 pmw between 1810s-1970s, and a steady increase from the 1980s onwards, from 2.65 pmw to 4.43pmw in the 2000s.

These numbers can only give us a rough estimate and an analysis of all instances will have to be performed to confirm whether the increase applies to the resultative type, the explanative type, or both. After manual sorting and analysis of the results, more can be said about the diachronic development and also about the conditions of use of the resultative type, e.g. are there certain adjectives or subject types that show a particular preference for the resultative ADJ + *enough* + *that*-clause construction? These issues are also addressed by Mindt (2011), but not specifically for the construction including *enough*.

If the initial hypothesis is substantiated, further avenues for research will include a more detailed investigation looking into reasons for such an increase; is it a signal of colloquialization; is the resultative construction used at the expense of other constructions; or does it increase by analogy with the more frequent explanative construction?

References

Francis, G., S. Hunston & E. Manning. 1998. *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

Mindt, I. 2011. *Adjective Complementation: An Empirical Analysis of Adjectives Followed by* that-*Clauses*. Amsterdam: Benjamins.

Multi-word organization names: A contrastive study of article use in English and German

Marianne Hundt, Elena Callegaro, Martin Volk & Johannes Graën (University of Zurich)

In both English and German, proper names are typically used without an article (Quirk et al. 1985, Biber et al. 1999, Huddleston and Pullum 2002, Engel 2004, Duden 2005). Against this general rule, multi-word organization names such as *the Department of Trade, Norwich Union, die Fraktion der Sozialdemokratischen Partei Europas* and *Ausschuss für Wirtschaft und Währung* provide an interesting case because they can be used to argue that common nouns and proper names are not two clearly distinct categories. In a previous study of article use in English organization names in three British English newspapers, Tse (2003) finds that certain types of premodifier (phrase names, acronyms, single-word proper nouns) make the omission of the article likely, whereas with other kinds of premodification (prepositional phrase, common noun phrase) multi-word organization names are highly likely to be used with an article. The first kind of premodification make the multi-word organization name structurally closer to proper nouns, the second one make it closer to common nouns. Her conclusion is that "these findings reinterpret in terms of gradience the 'classical' (although grossly oversimplified) assumption that common nouns require articles and proper names do not." Thus, multi-word organization names form a cline from those that are very similar to proper names and those that are more similar to common nouns.

In this paper, we use data from a parallel, sentence-aligned and part-of-speech tagged corpus of standard English and German (Europarl), a largely data-driven approach to data retrieval, and regression analysis. Overall, our results are compatible with Tse's findings: they confirm the proper-name to common-noun cline for English, albeit at a much lower incidence of multi-word organization names without articles; in the German data, instances of bare multi-word organization names are even rarer.

References

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

Duden. Die Grammatik. ^{7}2005. Hrsg. von der Dudenredaktion. Mannheim: Dudenverlag.

Engel, U. 2004. *Deutsche Grammatik*. München: Iudicium Verlag.

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

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

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.

A corpus-based analysis of lexical development in intermediate learners of English

Rolf Kreyer (Marburg)

The present study wants to explore the development of active vocabulary of German intermediate learners of English by analysing data from the Marburg corpus of Intermediate Learner English (MILE), which is currently being compiled at the University of Marburg (Kreyer, 2015). The corpus provides roughly 800,000 words of true-longitudinal data collected from German pupils during their final four years of secondary education (grades 9 to 12 in the German *Gymnasium*), thus providing an opportunity to analyse learner data on an intermediate level of proficiency, informants that have not been in the focus of learner corpus research so far.

The paper will zoom in on the development of vocabulary size, which will be tapped by different measures that gauge lexical variation and lexical richness in texts. Lexical variation will be explored by various versions of type-token-ratios (see Malvern et al. 2004 for a discussion). In addition, lexical richness will be analysed on the basis of the Lexical Freqency Profile (Laufer and Nation 1995), which assesses the proportions of more as opposed to less advanced lexis in a given text. More specifically, with the help of the AntWordProfiler 1.4.0 (Anthony 2012), the paper will describe the development of the proportions of vocabulary from different frequency levels (i.e. 1st 1000, 2nd 1000, other).

All in all, this work-in-progress report hopes to provide valuable corpus-based insights into the development of L2-lexis in intermediate learners of EFL.

References

Anthony, L. 2012. *AntWordProfiler (Version 1.4.0)* [Computer Software]. Tokyo, Japan: Waseda University. http://www.antlab.sci.waseda.ac.jp/.

Kreyer, R. 2015. The Marburg Corpus of Intermediate Learner English (MILE). In M. Callies & S. Götz eds. *Learner Corpora in Testing and Assessment*. Amsterdam: Benjamins, 13-34.

Laufer, B. & P. Nation. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, 16: 307-22.

Malvern, D. D., B. J. Richards, N. Chipere & P. Durán. 2004. *Lexical Diversity and Language Development: Quantification and Assessment.* Hampshire: Palgrave Macmillan.

Predicative adjectives and the correlation of speech rate and frequency

David Lorenz (Albert-Ludwigs-Universität Freiburg)

The present study explores the broad hypothesis that in rapid speech speakers use words that are more easily retrieved and processed. This idea follows from findings about the effects of speech rate and frequency.

The frequency of a word or string in a language has often been evoked as a determining factor in processes of both language production and processing (e.g. Bybee 2006, Diessel 2007). Production is faster for high-frequency words (Jurafsky et al. 2001), sequences (Arnon & Cohen Priva 2013) or syntactic patterns (Gahl & Garnsey 2004); high-frequency sequences are also processed more quickly (Arnon & Snider 2010). This is explained by the assumption that frequent items or structures are generally more predictable and that speakers and listeners "process the most likely events with the greatest ease" (Dell & Gordon 2003: 9). Rapid speech as such, however, is a potential source of processing difficulties in communication; speakers need to access lexical items faster when speaking above their usual rate. It is possible, then, that speakers mitigate these difficulties by reverting to more frequent words, which are more easily retrieved and processed.

This hypothesis is tested on conversation data from the Santa Barbara Corpus of Spoken American English (DuBois et al. 2000-2005), which provides audio recordings and time-aligned transcripts. Predicative adjectives were chosen as a testing ground because they are easy to identify, show a great deal of variation (e.g. the near-synonymous *great, awesome, excellent*, etc.) and tend to occur at the end of a phrase. The hypothesis predicts that higher speech rates lead to the use of adjectives that are generally more frequent.

Speech rate is operationalized as the rate of articulation (syllables per second) in an intonation unit, excluding the predicative adjective itself. The rate is measured as relative to the speaker's base line, i.e. their overall mean speech rate in the entire conversation. The frequency of each predicative adjective has been extracted from the Corpus of Contemporary American English (Davies 2008-), assuming that this very large corpus approximates the frequencies of items in the language at large.

Results indicate that there is indeed a positive correlation of speech rate and adjective frequency, which, however, is sensitive to other factors, especially a 'priming effect' where adjectives are repeated from prior discourse. This by and large confirms the hypothesis, but suggests that lexical retrieval in speech production is guided by many different aspects simultaneously. As the scope of this study is limited to predicative adjectives in American English, the question to what extent a correlation of speech rate and item frequency may hold in other contexts remains open.

On methodological grounds, the study shows, firstly, that psycholinguistic hypotheses mainly derived from experimental studies can be tested in natural speech data from corpora; and secondly, that the roles of speech rate and frequency in language production merit further attention.

References

Arnon, I. & U. Cohen Priva. 2013. More than words: the effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, 56:3: 349-71.

Arnon, I. & N. Snider. 2010. More than words: frequency effects for multi-word phrases. Journal of *Memory and Language*, 62: 67-82.

Bybee, J. 2006. From usage to grammar: the mind's response to repetition. *Language*, 82(4): 711-33.

Davies, M. 2008-. The Corpus of Contemporary American English: 450 million words, 1990-present. http://corpus.byu.edu/coca/.

Dell, G. S. & J. K. Gordon. 2003. Neighbors in the lexicon: friends or foes? In N. O. Schiller & A. S. Meyer eds. *Phonetics and Phonology in Language Comprehension and Production*. Berlin: Mouton de Gruyter, 9-38.

Diessel, H. 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25: 108-27.

Du Bois, J. W., W. L. Chafe, C. Meyer, S. A. Thompson, N. Marty & R. Englebretson. 2000-2005. *The Santa Barbara Corpus of Spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.

Gahl, S. & S. M. Garnsey. 2004. Knowledge of grammar, knowledge of usage: syntactic probabilities affect pronunciation variation. *Language*, 80(4): 748-75.

Jurafsky, D., A. Bell, M. Gregory & W. Raymond. 2001. Probabilistic relations between words: evidence from reduction in lexical production. In J. Bybee & P. Hopper eds. *Frequency and the Emergence of Linguistic Structure*. Amsterdam: Benjamins, 229-54.

COME *to*

Ilka Mindt (University of Paderborn)

The focus of this work-in-progress report is on forms of the verb COME followed by the word *to* as in

(1) +*Te seyd Barow sayd to me if he* ***com to*** *London* […] (Helsinki Middle English)

(2) *I hope, you did not* ***come to*** *rob me*? (Helsinki, Early Modern English)

(3) *That, I'd* ***come to*** *understand at last*. (COCA, Fiction)

In (1) COME is followed by the preposition *to*, whereas in (2) the function of *to* is to express a purpose. In (3) COME is followed by the *to*-infinitive marker.

The aim of the paper is to present a diachronic account of the development of the various options as given in (1) to (3). Special attention will be paid to semantic classes of verbs following COME *to* as in (2) and (3). A final point is to address the function of COME *to* V as a unit as well as the linguistic description of the word "to".

For this study, a range of corpora have been used: the Helsinki Corpus, the Corpus of Early English Correspondence Sampler, the Lampeter Corpus, the Corpus of Late Modern English Texts, the Old Bailey Corpus, the Corpus of Historical American English and the Corpus of Contemporary American English have been used.

Neology: Pinpointing the circumstances of coinage

Antoinette Renouf (Birmingham City University)

In studying neology in a large, diachronic text corpus, it is possible to observe new words emerging, evidencing a particular life-cycle, and then sometimes falling into disuse. But while a word in its full flush of use may be clearly observable, the circumstances of its birth are typically hazy.

Many linguists, however, require precise data on the nature of word coinage. In the EnerG neology project, for example, Schmid et al (2014) investigate the factors influencing the success and failure of neologisms, asking, "What are the circumstances under which new words are coined? What are the coiners' motives and aims?". The Oxford English Dictionary team (2014) strives "to pinpoint the genuine birthday of a word in English". Other linguists need information on provenance to evaluate the role of the hapax legomenon. In most NLP applications, hapax status, central to many studies (e.g. Baayen & Sproat, 1996) will not be automatically assigned if a word recurs, although there is a crucial difference between a word which recurs because one person borrows it directly from another; a word coined independently by unrelated people; and a word coined and shared by members of a team.

Nevertheless, the first occurrence of a new word in a written corpus is unlikely to represent its birth. Hohenhaus (2006) says "more often than not when we encounter words that are new to us these are clearly not completely new", while Lass (1980) writes that "we can never observe the exact moment when a change begins (except by accident)". Mair (1996) remarks that "nothing really follows from the fact that an early written attestation of 'something new' happens to be included in Frown or FLOB".

In this work-in-progress report, we examine the issue of pin-pointing neologisms at birth in written newstext, and seek to offer some solutions to the problem facing linguists. We base our study on a diachronic corpus of 1.4 billion-words of UK *Guardian* and *Independent* newspapers, currently from 1984-2014. This corpus is stored, processed and presented for analysis by the WebCorp Linguist's Search Engine (begun in 2006) (http://wse1.webcorp.org.uk/home/). Candidate neologisms are also proposed by APRIL software (1997-2000) (http://rdues.bcu.ac.uk/aprdemo/). We examine the contexts which surround words coined both by the coiner him/herself, and by writers acknowledging the origin of another's coinage. In the latter case, we are finding neologisms accompanied by explicit flags of attribution, such as *coined* or *neologism*, often supported by orthographic signals of novelty (Renouf & Bauer, 2001). Yet these signals rarely fix the birth of a neologism at a precise point in time, a fact which we shall try to explain. Where the first occurrence of a word in a diachronic corpus represents its birth, we differentiate between deliberate coinages, to which coiners may wish to lay claim, and on-the-fly coinages, which may be sub-conscious and only intermittently flagged. We also examine cases where a neologism is loosely flagged by *uttered*, or *cited*, or where a lexical item which is not a neologism signalled as *coined*.

In these ways, we hope to clarify the situation regarding the birth of coinages, and develop search guidelines to facilitate the automated discovery of provenance and birth date of word coinage in written text.

Acknowledgements

Thanks are due to the Engineering and Physical Science Research Council, under grant nos. EP/E001300/1 and GR/L08243/01. The WebCorp Linguist's Search Engine and APRIL systems were developed within the Research and Development Unit for English Studies at Birmingham City University.

References

Baayen, R. H. & R. Sproat. 1996. Estimating lexical priors for low-frequency morphologically ambiguous forms. *Computational Linguistics*, 22: 162.

Hohenhaus, P. 2006. Bouncebackability: A Web-as-corpus-based case study of a new formation, its interpretation, generalization/spread and subsequent decline. *SKASE Journal of Theoretical Linguistics*, 3(2):17-27.

Lass, R. 1980. *On Explaining Language Change*. Cambridge: CUP.

Mair, C. 1996. Parallel corpora: A real-time approach to the study of language change in progress. In M. Ljung (ed.) *Corpus-based Studies in English*. Rodopi: Amsterdam Atlanta. 195-209.

Oxford Dictionaries. 2014. Tracing the birth of words: from 'open' to 'heffalump'. At http://blog.oxforddictionaries.com/2012/04/tracing-the-birth-of-words/.

Renouf, A. & L. Bauer. 2001. Contextual clues to word-meaning. *International Journal of Corpus Linguistics*, 5(2): 231-258.

Schmid, H.-J. 2014. Words, words, words. http://www.anglistik.uni-muenchen.de/abteilungen/sprachwissenschaft/research/research_projects1/energ/index.html.

Desert Island Discs and recent language change

Nick Smith (University of Leicester)

This paper discusses the rationale, design and early findings from a small corpus based on the popular BBC radio programme, Desert Island Discs. The main impetus for the study comes from recognition that despite significant advances in the last two decades, research on recent language change in English has depended for the most part on written, published resources. On the one hand, corpora such as the Brown family and COHA have helped to uncover dramatic shifts in the frequency and use of different constructions in British and American English, for example, in the progressive, the core modals and semi-modals, noun-noun sequences, and complementation patterns (Hundt and Mair 1999, Leech et al. 2009, Rudanko 2011, Aarts et al. 2013). These corpora have, moreover, given support to factors and theories of change, notably grammaticalization and colloquialization. On the other hand, without a fuller representation of spoken corpora in the recent change 'landscape', it is difficult to move beyond or add flesh to claims based largely on written English. One step towards redressing the imbalance has been the Diachronic Corpus of Parsed Spoken English (DCPSE), which contains speech recordings from the 1960s to the 1990s. The present paper

represents a further step, in that it tracks a single, spoken genre of British English across six decades.

The study draws on a highly promising, but so far underutilized, resource for linguistic analysis – the multifarious collection of 20th- and 21st-century radio and television programmes in the BBC Archive. Desert Island Discs (DID) is particularly promising in this respect. DID is an early form of radio broadcast discussion or 'chat show' (Castell 1999: 392, Tolson 2006), in which celebrities discuss with an interviewer their lives, past and present, and comment on their favourite pieces of music. The rationale for using DID in this study includes, firstly, the longevity and consistency of the programme: it has been broadcast almost continuously each week from 1942 to the present day, with minimal changes in the format, thus making it possible to track linguistic changes reasonably precisely. Secondly, DID has always been pitched as a lighter kind of programme, involving convivial, dialogic exchanges, and a reasonably broad social spectrum of interviewees (Magee 2012). It thus potentially affords valuable insights on colloquial developments in 20th-century English. Another, practical, advantage of DID is that it can be downloaded, unlike most of the (streamed) material currently available on the Archive.

The paper will describe and evaluate efforts to sample the Archive material in a principled, consistent way, to produce a mini-corpus that is conducive to investigating change in spoken broadcast English. Issues in the design and creation of the corpus, and how they have been tackled, will be discussed, namely:

i) the periodization or date-sampling of the corpus;
ii) the selection and social representativeness of the participants: initially these will include native speakers of British English only;
iii) change over time in the role of the interviewer;
iv) transcription and coding conventions.

Thus it is hoped that the present study, although focused on a certain kind of English within a single genre, will help to establish some principles and directions for the incorporation of sound archives in corpus linguistic research, particularly studies focused on recent language change.

References

Aarts, B., J. Close, G. Leech & S. Wallis eds. 2013. *The Verb Phrase in English: Investigating Recent Language Change with Corpora*. Cambridge: Cambridge University Press.

Castell, S. 1999. Phone-ins and chat shows. In P. Childs & M. Storry eds. *Encyclopaedia of Contemporary British Culture*. London: Routledge, 392-93.

Hundt, M. & C. Mair 1999. "Agile" and "uptight" genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4(2): 221-42.

Leech, G., M. Hundt, C. Mair & N. Smith 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Magee, S. 2012. *Desert Island Discs: 70 Years of Castaways*. London: Random House.

Rudanko, J. 2011. *Changes in Complementation in British and American English: Corpus-Based Studies on Non-Finite Complements in Recent English*. Basingstoke: Palgrave.

Tolson, A. 2006. *Media Talk: Spoken Discourse on TV and Radio*. Edinburgh: Edinburgh University Press.

I was stood there ironing! A corpus-based study on the distribution of pseudo-passives with *sat* and *stood* in British and American English

Ulrike Stange (Johannes Gutenberg-Universität, Mainz)

The present paper is concerned with pseudo-passives containing *sat* and *stood*, which are functional equivalents of the corresponding standard progressive constructions with *sitting* and *standing*:

(1) (a) He **was sat** there and he was going brr brr brr brr brr. (*BNC* KCE: 4738)
(b) Last night [pause] I **was sitting** er [sic] what was I watching? (*BNC* KCJ: 216)

(2) (a) Well when you**'re stood** there you can see the flames. (*BNC* KD2: 303)
(b) Wondered what I **was standing** on there it's a bit of squashed carrot. (*BNC* KB8: 11455)

Utterances (1a) and (2a) look like passives sentences in that the auxiliary *be* is followed by a past participle, but their function is competing with the standard forms as they appear in (1b) and (2b) (*be* + present participle). As they are active in semantics and passive in form they are termed *pseudo-passives*. In Standard English, these constructions are deemed incorrect unless they are proper passive constructions, i.e. semantically and formally passive (cf. Wood 1962[1981]: 235, 253), for "they could only mean that the person in question was placed there by someone else" (Wood 1962[1981]: 196). Furthermore, pseudo-passives are a feature of regional usage and "very common, even amongst well-spoken people" (Wood 1962[1981]: 235).

Traditionally, pseudo-passive constructions with *sat* and *stood* have been associated with British English dialects, particularly with those of the North and the Southwest (cf. Klemola 1999, 2002, Cheshire et al. 1989, 1993). However, recent studies have shown that this dialectal feature is not only spreading within England (cf. Stange forthc.), it is also becoming more frequent in British newspapers (Rohdenburg/Schlüter 2009). Furthermore, the use of pseudo-passives with *sat* and *stood* is not restricted to the British Isles, but found in a number of Englishes around the world, e.g. in Newfoundland English and in Rural and Urban African American Vernacular English (eWAVE, Kortmann/Lunkenheimer 2013).

The present paper explores the distribution of *be sat* and *be stood* across spoken and written registers in British and in American English, using data drawn from the *British National Corpus (BNC)* and the *Corpus of Contemporary American English (COCA)*. First results have shown that these constructions are significantly more frequently found in spoken data in general and in British data in particular – which is to be expected considering that pseudo-passives are a dialectal feature and that they originate in England. Where *be sat* and *be stood* with progressive meaning occurred in written data, they had their source in spoken language more often than not (as in interviews, dialogues, etc.). While Rohdenburg/Schlüter (2009) found no pseudo-passives with *sat* and *stood* in their American newspaper data, the present study is able to show that they are being used in American English in spoken discourse and in some written genres (fiction, magazines). Computing frequencies and testing them for statistical significance, this paper will present a survey of the distribution of pseudo-passives

with *sat* and *stood* across spoken and written registers, highlighting British-American as well as register-based differences and similarities.

References

Davies, M. 2008-. The Corpus of Contemporary American English: 450 million words, 1990-present. http://corpus.byu.edu/coca/.

Davies, M. 2004-. BYU-BNC. (Based on the British National Corpus, Oxford University Press). http://corpus.byu.edu/bnc/.

Cheshire, J., V. Edwards, & P. White. 1989. Urban British dialect grammar: the question of dialect levelling. *English World-Wide* 10(2): 185-225.

Cheshire, J., V. Edwards, & P. White. 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.

Klemola, J. 1999. 'Still sat in your car?' Pseudo-passives with sat and stood and the history of non-standard varieties of English English. *Sociolinguistica* 13: 129-40.

Klemola, J. 2002. Continuity and change in dialect morphosyntax. In D. Kastovsky, G. Kaltenböck & S. Reichl eds. *Anglistentag 2001 Wien*, Trier: Wissenschaftlicher Verlag, 17-56.

Kortmann, B. & K. Lunkenheimer eds. 2013. *The Electronic World Atlas of Varieties of English [eWAVE]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://www.ewave-atlas.org/. Accessed 2014-12-14.

Rohdenburg, G. & J. Schlüter 2009. New departures. In G. Rohdenburg & J. Schlüter eds. *One Language, Two Grammars*, Cambridge: CUP, 364-423.

Stange, U. forthcoming. I was sat there talking all night: A corpus-based study on factors governing intra-dialectal variation in British English. English Language and Linguistics. Special Issue: *Support Strategies in Language Variation and Change*.

Wood, Frederick T. 1962 [1981]. *Current English Usage*. Hamburg: PMV.

Morphosyntactic creativity in relativization processes of second-language varieties

Cristina Suárez-Gómez (University of the Balearic Islands)

This paper examines morphosyntactic creativity in indigenized second-language varieties of English, analyzing in detail innovative traits found in relative clauses and processes of relativization. Basing the analysis on data from the *International Corpus of English* (http://ice-corpora.net/), it explores adnominal relative clauses (e.g. "I can't make it up with uhm a person *to whom I seldom speak English*" <ICE-HKE S1A-033#16:1:A>) in the spoken language of the varieties found in Hong-Kong, India and Singapore and compares them with British English.

One of the most relevant findings of this large-scale study affects the different distribution of relative words in comparison with native varieties of English. This divergence from the norm is usually related to a different usage of the relative words, both in frequency of use and difference in stylistic and/or structural preferences, but it does not generally affect grammaticality conditions (see also Huber 2012: 240). Such divergent structures in L2 varieties of English (e.g. "They've reached a real good status to over the men in each and every field *that they have participated*" <ICE-IND:S1A-011#45:1:F> illustrating 'preposition

chopping') are usually catalogued either as performance or planning errors or as linguistic innovative features, and the difficulty in distinguishing both has been a subject of debate (Bamgbose 1998: 1-2; Van Rooy 2011: 192); hence these structures have generally been omitted from quantitative analyses. This practice will be challenged in the present paper, because quantitative analyses actually demonstrate that they are not isolated cases in the varieties under study and they may constitute features that support linguistic nativization. They often illustrate patterns that recur in different non-native varieties and may thus represent features characteristic of contact-induced processes, whereas on other occasions they seem to be local features which over time may become consolidated as local norms, as has been already demonstrated, for example, in the verbal complementation system of different New Englishes (Schneider 2004; Mukherjee & Hoffmann 2006; Mukherjee & Gries 2009; Van Rooy & Terblanche 2009) and in other linguistic levels, as is the case of phonology (Gut 2007; Hung 2009).

The aim of this paper is to analyze morphosyntactic innovations in the realm of relativization which might constitute instances of incipient change and can thus be claimed as nativized features. It will examine in detail the potential factors determining the innovations, whether they are attributable to the relevant L1s, to the standard variety of the target language or to both of these, or whether they are the result of the development of the contact varieties themselves.

References

Bamgbose, A. 1998. Torn between the norms: Innovations in world Englishes. *World Englishes* 17(1): 1-14.

Gut, U. 2007. First language influence and final consonant clusters in the new Englishes of Singapore and Nigeria. *World Englishes* 26(3): 346-359.

Huber, M. 2012. Syntactic and variational complexity in British and Ghanaian English. Relative clause formation in the written parts of the International Corpus of English. In B. Kortmann & B. Szmrecyani (eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: Walter de Gruyter, 218-242.

Hung, T. T.N. 2009. Innovation in second language phonology. In L. Siebers & T. Hoffmann (eds.), *World Englishes–Problems, Properties and Prospects*. Amsterdam & Philadelphia: John Benjamins, 227–38.

International Corpus of English. 1990-. Available through the ICE Project. http://ice-corpora.net/ice.

Mukherjee, J. & S. Hoffmann. 2006. Describing verb-complementational profiles of new Englishes: a pilot study of Indian English. *English World-Wide* 27(2): 147-173.

Mukherjee, J. & S. Gries. 2009. Verb-construction associations in the International Corpus of English. *English World-Wide* 30(1): 27-51.

Schneider, E. 2004. How to trace structural nativization: particle verbs in world Englishes. *World Englishes* 23(2): 227-249.

Van Rooy, B. 2011. A principled distinction between error and conventionalized innovation in African Englishes. In J. Mukherjee & M. Hundt (eds.), *Exploring Second-Language Varieties of English and Learner English. Bridging a Paradigm Gap.* Amsterdam & Philadelphia: John Benjamins, 189-207.

Van Rooy, B. & L. Terblanche. 2009. Complementation patterns in causative verbs across varieties of English. Paper presented at *Corpus Linguistics 2009*, Liverpool.

Developing a corpus of Early Modern English military writing

Jukka Tuominen (University of Tampere)

More than a decade after Peter Burke (2004: 129–130) noted the surprising lack of research on the linguistic practices of early modern soldiers and officers, the field remains largely uncharted. Although some recent studies have considered the multilingual contexts of warfare in the late Middle Ages (Curry et al. 2010), from the eighteenth century to the present (e.g. Footitt and Kelly 2012), and with a focus on German and French military men (Glück and Häberlein 2014), the compilation and analysis of British sources from the early modern period is a task that still awaits scholarly attention.

The military sphere is particularly interesting as a potentially rich setting for language contact. Since early modern armies could include units or individual soldiers from many nationalities, communication both among the troops and with civilians and captives would often have necessitated a level of multilingual interaction. Moreover, warfare, like other professional domains, involved the exchange and dissemination of information across political and linguistic boundaries. The military elite needed to keep informed of developments abroad, and played a key role in transmitting knowledge of foreign innovations to the wider community.

The present project is part of a broader study of multilingual discourse practices in different domains of Early Modern English professional writing. The primary aim of compiling a new corpus of English military texts is to provide a representative dataset for comparisons across domains and over time, but also for linking observed multilingual practices with language-external sociolinguistic variables, including the educational background of the writer and the intended readership of the text.

The initial stage of corpus compilation focuses on printed texts, utilizing the SGML/XML-encoded materials made available by the Early English Books Online Text Creation Partnership (EEBO-TCP). The selection of texts, made in consultation with a military historian, aims to reflect the variety of military texts published for professional and non-professional readers. Salient genres of early modern military publishing include treatises on the philosophical and legal aspects of war, technical manuals and guidebooks, accounts of battles and campaigns, memoirs, and military history (Donagan 1995; Lawrence 2009).

The intention is to eventually develop the corpus further in terms of both the temporal limits and the number and range of texts included. This will entail extending the coverage into the Late Modern English period and complementing the published works with private and documentary materials such as personal letters, diaries and notebooks, as well as the records of military units.

References

Burke, P. 2004. *Languages and Communities in Early Modern Europe*. The 2002 Wiles Lectures, Queen's University, Belfast. Cambridge: Cambridge University Press.

Curry, A., A. Bell, A. Chapman, A. King & D. Simpkin. 2010. Languages in the military profession in later medieval England. In R. Ingham ed. *The Anglo-Norman Language and its Contexts*. Woodbridge: York Medieval Press in association with The Boydell Press. 74-93.
Donagan, B. 1995. Halcyon days and the literature of war: England's military education before 1642. *Past & Present;* 147: 65-100.
Footitt, H. & M. Kelly eds. 2012. *Languages and the Military: Alliances, Occupation and Peace Building*. Palgrave Studies in Languages at War. Basingstoke: Palgrave Macmillan.
Glück, H. & M. Häberlein eds. 2014. *Militär und Mehrsprachigkeit im neuzeitlichen Europa*. Fremdsprachen in Geschichte und Gegenwart, 14. Wiesbaden: Harrassowitz Verlag.
Lawrence, D. R. 2009. *The Complete Soldier: Military Books and Military Culture in Early Stuart England, 1603–1645*. History of Warfare, 53. Leiden and Boston, MA: Brill.

Negotiating expectations and the identification of ideology in LModE historiographical texts

Sebastian Wagner (University of Duisburg-Essen)

The works of historians shape their readers' view on the past. Even though it is well known that the historiographical text reflects its author's subjective political, social, cultural, religious and educational stances (Warren 1998: 2), the linguistic encoding of ideologies within historical discourse is, thus far, still a rather under-researched area (in spite of the pioneering contributions of Martin & Wodak 2003, Coffin 2006 and Verschueren 2012).

One way of approaching these underlying ideological meanings is by examining the ways in which the historian (re-)evaluates past events. Combining Thompson and Hunston's (2000: 8) notion of ideologies as being "essentially sets of values" with Verschueren's (2012: 12) understanding of ideological meaning as "rarely questioned" and perceived as "commonsensical" provides a preliminary working definition of *ideology* on the basis of which the historical writer's expression of evaluative meaning within the boundaries of the particular value-system prevailing in a society can be analysed.

Embracing a dialogistic perspective on evaluation, this contribution seeks to focus on the ways in which an imagined reader is positioned with regard to the presented proposition. Evaluating along the parameter of EXPECTEDNESS (Bednarek 2006: 44; cf. also Thompson & Hunston 2003: 23f.) the author construes a putative addressee who, for the most part, is expected to share his/her value position. Notions of unexpectedness, conversely, are signalled as COUNTER EXPECTATIONS, CONTRAST and negation/denial (Bednarek 2006: 48, Martin & White 2005: 118). The present study tries to show that by analysing the use of adversative and concessive expressions (e.g. *but, yet, although, however, still, even though*) in relation to their immediate co-text one might in turn be able to disclose indications for the interpretation of preceding (unevaluated/stance-neutral) propositions. These propositions are typically realized through "unmodalized positive declaratives", which are believed to implicitly encourage the reader to align with the author's perspective (Coffin 2009: 143). Martin & White (2005: 121) suggest that counters (contrasts), in a similar manner, function to construe the writer as aligning with the reader. Thus, if the writer articulates his/her surprise about an

unexpected situation the reader is inclined to subscribe to these clearly marked unexpected propositions.

In historical writings these counters/contrasts frequently serve as a means to evaluate, for instance, specific circumstances or behaviours on the basis of a perceived commonsensical set of social norms and values (which is assumed to be at least partly informed by ideology).

> Such disorders are the natural effects of religious tyranny, **but** the rage of the Donatists was inflamed by a frenzy of a very extraordinary kind (Gibbon 1782)

In the main clause, Gibbon presents the disorders as an expected ('natural') phenomenon resulting from religious tyranny. The reader is inclined to conceive the proposition, which noticeably echoes Gibbon's anti-religious stance, as the 'norm'. The evaluation by counter-expectation, as initiated by the but-clause, results from the endeavour to make the reader subscribe to a disproportionally (i.e. unexpectedly) vile demeanour on the part of the Donatists. In this way, the Christian secession movement is evaluated even more negatively.

The extract illustrates that historians tend to mark any notion of deviation from the perceived norm as unexpected. By doing so, they constrain alternative voices and deviant ideologically informed value positions.

The analysis is conducted using a small pilot corpus of historical writing that is part of a larger corpus project covering ca. 1200 years of British historiography. The pilot corpus comprises selected works of British 18th and 19th-century historians. The time period in focus thus not only covers philosophical historiography (still heavily influenced by the ideas of Enlightenment) but likewise includes the representatives of the emerging professionalized, scientific school of historiography (cf. Stuchtey 1999: 37).

References

Bednarek, M. 2006. *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. London/New York: Continuum.

Coffin, C. 2009. *Historical Discourse: The Language of Time, Cause and Evaluation.* London/New York: Continuum.

Hunston, S. & G. Thompson. 2003. *Evaluation in Text. Authorial Stance and the Construction of Discourse*. Oxford: OUP.

Martin, J. R. & P. R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Basingstoke: Palgrave.

Martin, J. R. & R. Wodak. 2003. *Re/reading the Past: Critical and Functional Perspectives on Time and Value*. Amsterdam: Benjamins.

Stuchtey, B. 1999. Literature, liberty, and life of the nation. British historiography from Macaulay to Trevelyan. In S. Berger et al eds. *Writing National Histories. Western Europe since 1800.* London/New York: Routledge, 30-46.

Verschueren, J. 2012. *Ideology in Language Use: Pragmatic Guidelines for Empirical Research.* Cambridge: Cambridge University Press.

Warren, J. 1998. *The Past and Its Presenters: An Introduction to Issues in Historiography*. London: Hodder Education.

A corpus-based analysis of surveillance discourses in the *Daily Telegraph*, the *Times* and academic research articles

Viola Wiegand (University of Nottingham)

In 2013 a worldwide public discussion on privacy and surveillance was triggered by the leak of confidential data from the US National Security Agency, alleging that the agency has been collecting massive amounts of emails and call metadata (Black, 2013). Given the rapid worldwide expansion of surveillance measures since the September 11 terror attacks in 2001 (Lyon, 2004), *surveillance* is arguably becoming a 'cultural keyword' (Williams, 1983) and appears to gain great social significance. While recently security and surveillance discourses have gradually attracted more research interest (e.g. Barnard-Wills, 2011; MacDonald & Hunter, 2013a, 2013b; MacDonald et al., 2013), still only relatively few studies have examined public discourses of surveillance from an explicitly linguistic perspective. Branum and Charteris-Black (2015) study the media reception of the Edward Snowden case, but focus mainly on this particular event and its ideological news coverage using keyword analysis. Indeed, a corpus linguistic approach allows us to systematically analyse "the representation of social issues, global events or groups in society" (Mahlberg, 2014: 220) by means of examining linguistic patterns. While some corpus linguists adopt concepts from Critical Discourse Analysis (e.g. Baker, 2006; Baker et al., 2008), corpus linguistic investigations can also draw on theoretical models from other disciplines. For instance, McEnery (2009) has shown that 'key keywords' can be employed to test the validity of a sociological theory in a corpus. Similarly, the present study proposes a corpus-assisted approach to the discursive construction of surveillance in newspaper articles by comparing key keywords to a sociological model of surveillance 'frames' (Barnard-Wills, 2011). The research questions addressed at this stage of the project are: First, how can we use the term *surveillance* to collect newspaper corpora representing surveillance discourses? Second, to what extent can key keywords of such specialised surveillance news corpora be mapped onto Barnard-Wills' (2011) surveillance frames? And finally, how can we describe the meaning of surveillance in these corpora? Patterns around the discourses of surveillance are analysed using *WordSmith Tools* (Scott, 2012). The data studied compromises a pilot corpus of surveillance articles from 2001-2005 in the *Daily Telegraph* and a sub-section of the Times Digital Archive. Findings to date indicate that surveillance discourses in the Daily Telegraph Surveillance Corpus match the surveillance frames with slight revisions. A strong discursive distinction between 'them' − terrorists − and 'us' − a union of the USA and Britain − suggests that the discourses reflect the sampling period of 2001-2005. Concordance lines reveal various meanings of *surveillance* including those of terrorist prosecution, crime prevention and meanings in the contexts of epidemic, economic and pedagogic monitoring. It is expected that the analysis of the *Times* articles will shed more light on potential diachronic changes in the representation of surveillance before and after the September 11 terror attacks. An additional dimension of the project will contrast the newspaper findings with surveillance discourses in the academic journal *Surveillance & Society*, in order to examine differences, similarities and potential instances of intertextuality between these corpora.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

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

Barnard-Wills, D. 2011. UK news media discourses of surveillance. *The Sociological Quarterly,* 52(4), 548-67.

Black, I. 2013. NSA spying scandal: what we have learned. *The Guardian,* June 10. http://www.theguardian.com/world/2013/jun/10/nsa-spying-scandal-what-we-have-learned

Branum, J., & J. Charteris-Black 2015. The Edward Snowden affair: a corpus study of the British press. *Discourse & Communication,* 9(2), 199-220.

Lyon, D. 2004. Globalizing surveillance: comparative and sociological perspectives. *International Sociology,* 19(2), 135-49.

MacDonald, M., & D. Hunter. 2013a. The discourse of Olympic security: London 2012. *Discourse & Society,* 24(1), 66-88.

MacDonald, M., & D. Hunter. 2013b. Security, population and governmentality: UK counter-terrorism discourse (2007-2011). *Critical Approaches to Discourse Analysis Across Disciplines,* 7(1): 123-40.

MacDonald, M., D. Hunter & J. O'Regan 2013. Citizenship, community, and counter-terrorism: UK security discourse, 2001-2011. *Journal of Language and Politics,* 12(3): 445-73.

Mahlberg, M. 2014. Corpus linguistics and discourse analysis. In K. P. Schneider & A. Barron eds. *Pragmatics of Discourse*. Berlin: De Gruyter Mouton, 215-38.

McEnery, T. 2009. Keywords and moral panics: Mary Whitehouse and media censorship. In D. Archer ed. *What's in a Word-list?: Investigating Word Frequency and Keyword Extraction.* Farnham: Ashgate, 93-124.

Scott, M. 2012. *WordSmith Tools 6*. Liverpool: Lexical Analysis Software.

Williams, R. 1983. *Keywords: A Vocabulary of Culture and Society*,2nd ed. London: Fontana.

A permutation-based perspective on morphosyntactic differences between registers

Christoph Wolk (Justus-Liebig-Universität Gießen)

Data-driven analysis of lexical sequences in registers has received considerable attention in recent years (e.g. Gries, Newman & Shaoul 2011). Fewer studies have focused on a more abstract and/or syntactic level, where the popular approaches, such as Douglas Biber's multi-dimensional analysis (1988), rely on manually pre-specified features. In this talk, I will present the first results of combining the permutation-based part-of-speech n-gram analysis pioneered by (Nerbonne & Wiersma 2006) with a register-oriented perspective.

Nerbonne & Wiersma (2006) proposed a simple measure for determining the syntactic distance between two corpora. Their method first constructs and counts all sequences of length *n* in a part-of-speech annotated corpus. For example, the sentence "What_DDQ would_VM you_PPY do_VDI next_MD" (ICE-NZ S1B-073) yields, for *n=2* (i.e. *bigrams*), the following sequences: DDQ.VM, VM.PPY, PPY.VDI, VDI.MD. The counts are then normalized, and the result is compared to many runs of the same procedure on corpora that are randomly resampled from the original material. This allows a statistical analysis of the

individual n-gram distributions that is robust against some dispersion issues: sequences that only appear in very few texts are likely to be highly variable between runs, and thus fail to achieve statistical significance (see also Lijffijt, Säily & Nevalainen 2012). In addition to identifying robust individual differences between corpora, the total distance between them can also be quantified. While early uses of this method compared only two corpora, extensions that permit application to multiple corpora at the same time were added later, mainly for dialectometric purposes. The first was Sanders (2010), who used a variant to analyze a Swedish dialect corpus. Wolk (2014) introduced the reliability matrix, which allows an analysis when all corpora are permuted at once. The application to the *Freiburg Corpus of English Dialects* Sampler showed that this metric outperformed the frequency-based analysis and yielded results comparable to manually-selected and extracted features of dialect morphosyntax.

I will present the first results of this method on the recently released part-of-speech annotated versions of several components of the *International Corpus of English* (ICE, Greenbaum 1996), focusing on variation within the 32 registers. Overall, the method yields plausible results, both concerning the overall relations between the registers and the n-grams that emerge as distinctive. The method correctly distinguishes spoken and written language, with some expected deviations (e.g. *business/social letters* and *novels & stories* grouping with spoken instead of written material). The most distinctive patterns (e.g. *possessive pronoun + adverb*, distinctive for the two letters registers, such as *yours sincerely/faithfully* etc. or various combinations involving interjections, distinctive for spoken material) are unsurprising, but highly plausible. Overall, the results bear a strong resemblance to those reported in Gries, Newman & Shaoul (2011), indicating an overall convergence between lexical and part-of-speech n-grams.

References

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

Greenbaum, S. 1996. *Comparing English Worldwide: the International Corpus of English*. Oxford, New York: Clarendon Press.

Gries, S. Th., J. Newman & C. Shaoul. 2011. N-Grams and the clustering of registers. *ELR Journal* 5 (1).

Lijffijt, J., T. Säily & T. Nevalainen. 2012. CEECing the baseline: lexical stability and significant change in a historical corpus. In J. Tyrkkö, M. Kilpiö, T. Nevalainen & M. Rissanen eds. *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. Studies in Variation, Contacts and Change in English 10. http://www.helsinki.fi/varieng/series/volumes/10/lijffijt_saily_nevalainen/.

Nerbonne, J. & W. Wiersma. 2006. A measure of aggregate syntactic distance. In J. Nerbonne & E. 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.

Sanders, N. C. 2010. *A Statistical Method for Syntactic Dialectometry*. PhD thesis, Indiana University Bloomington. http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED523477.

Wolk, C. 2014. *Integrating Aggregational and Probabilistic Approaches to Language Variation*. PhD thesis, Freiburg: University of Freiburg.

The Eighteenth-Century English Phonology Database: The phonology of lexical sets in Eighteenth-Century English

Nuria Yáñez-Bouza (University of Vigo / University of Manchester)
Joan C. Beal (University of Sheffield)
Ranjan Sen (University of Sheffield)
Christine Wallis (University of Sheffield / Newcastle University)

This presentation reports on work in progress towards a database of Eighteenth-Century English Phonology (ECEP), a sister project to the Eighteenth-Century English Grammars database. ECEP will be a searchable database which can serve as a source bank for quantitative and qualitative studies, thereby meeting the demands of the growing research community in historical phonology and dialectology in particular.

The proposed database is being developed in three steps:

(i) Selection of primary sources that contain information about the phonology of 18C English, with a focus on pronouncing dictionaries. To date we have consulted Kenrick (1773), Perry (1775), Sheridan (1780), Walker (1791), Jones (1798) and Johnston (1764).

(ii) Data input and annotation. The database has been built as a relational database, constructed with integrated tables and a variety of 'fields' and 'forms'. The data are being systematically annotated and thematically grouped in three major categories: phonology-related data, source-related metadata and author-related metadata.

(ii-a) Phonology-related data. Each 18C pronunciation source is examined for illustrative examples of Wells's Standard Lexical Sets of vocalic variants (1982: 127–68), which include c. 30 sets and c. 60 subsets, adding up to over 1,700 individual keywords. To these we have added supplementary sets of consonantal variants. Each keyword is transcribed into unicode IPA, and is further annotated by lexical set and subset. We plan to add as well data about frequency of the lexical items in 18C English, and metalinguistic comments. The use of these sets and their associated keywords is standard practice in studies of variation and change in English; including the full range of example words allows for differences in lexical distribution between the primary sources, and between these and the contemporary accents described by Wells. Thus, a researcher interested in the distribution of words in Wells's PRICE and CHOICE sets would be able to find how each of the example words from these sets was transcribed in each of the 18C sources documented in the database, and how phonological variants were perceived at the time in the context of the standardisation of English (e.g. correct, dialectal, vulgar, etc.).

(ii-b) Source-related metadata, such as title, year of publication, editions, place of publication, imprint information, physical description, target audience.

(ii-c) Author-related metadata, such as name, life-dates, gender, social class, birthplace, place of residence and occupation.

(iii) Web-based application. After completion of the data input and verification processes, the database will be migrated to a web-based application hosted on the Humanities Research

Institute website using client-side HTML and Javascript and server-side PHP and MySQL. The proposed digital resource will be made freely available in a visually and technically user-friendly interface. The interface will display two layouts – browse and search – and will also offer a download function in CVS file format.

References

Eighteenth-Century English Grammars database. http://www.manchester.ac.uk/ECEG

Eighteenth-Century English Phonology database. http://hridigital.shef.ac.uk/eighteenth-century-english-phonology.

Jones, Stephen. 1798. *Sherdan Improved: A General Pronouncing and Explanatory Dictionary of the English Language*, 3rd Edition. London

Johnston, William. 1764. *A Pronouncing and Spelling Dictionary*. London.

Kenrick, William. 1773. *A New Dictionary of the English Language*. London.

Perry, William. 1775. *The Royal Standard English Dictionary*. Edinburgh.

Sheridan, Thomas. 1780. *A General Dictionary of the English Language*. London.

Walker, John. 1791. *A Critical Pronouncing Dictionary and Expositor of the English Language*. London.

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

The use of *although*-clauses in novice academic writing: Effects of nativeness and genre

Ekaterina Zaytseva (University of Bremen)

This study focuses on adverbial clauses of concession introduced by the subordinator *although* in the writing of two groups of novice academic writers: German advanced learners of English as an L2 and English native speakers. It explores variation in the use of concessive clauses with a focus on their discourse-pragmatic functions and sentence positioning. The study combines pedagogical and variationist approaches to learner language and analyses possible effects of factors like genre and nativeness, along with information status and discourse-pragmatic functions on the use of *although*-clauses by these writers.

Concessive clauses contain information that stands in opposition to the statement of the main clause (Biber et al. 1999: 779) and that is viewed as "wrong or surprising" within the context of a sentence (Siepmann et al. 2008: 357). In English native academic prose, they often serve to assist writers in argument development when discussing and evaluating previous research and/or presenting results of their own studies (cf. Gillet 2014, Morley 2014). In terms of their structural characteristics, concessives were found to occur in all three sentence positions, i.e. initial, final, and medial, with a strong preference for the former two positions (Biber et al. 1999). Previous comparative studies of L2 and L1 research papers have additionally revealed that although it is found in L1, medial position is not attested in L2 academic writing (see Kerz 2011). Furthermore, L2 and L1 argumentative essays were found to differ in terms of general frequency, meaning, and preferences in sentence positioning of concessive markers in writing (e.g. Crew 1990, Granger & Tyson 1996, Kerz 2011, Wagner 2011).

It still remains unclear, however, whether genre influences the use of "although"-concessives in both L1 and L2 novice academic writing and to what extent findings regarding the effect of native language, obtained separately for research papers and argumentative essays, could be generalized. The present contribution examines these issues and addresses the following research questions:

1. Which factors affect the use of *although*-clauses in L2 and L1 novice academic writing?
2. To what extent is variation in the use of *although*-clauses in novice academic writing influenced by genre as one plausible variable?
3. Are variation effects induced by writers' L1 consistent across two genres (i.e. research papers and argumentative essays)?

In the majority of learner corpus studies to date, learners' language in the written mode has been primarily investigated in either argumentative essays or, only recently, in research papers where it has been mainly approached from a pedagogical perspective, i.e. focusing on L2 and L1 differences and interpreting those in the light of learners' native language, without taking into account other factors such as genre (see, however, Ädel 2008, Paquot et al. 2011, Römer 2009). Meanwhile, the combination of pedagogical and variationist approaches considering a variety of factors may provide further explanations of advanced learners' choices in writing and allows for a more encompassing view of L2 novice academic writers as active language users, rather than only EFL learners.

The study draws on four comparable sections of L2 and L1 corpora of novice academic written English: the *International Corpus of Learner English* (ICLE; Granger et al.2009), the *Corpus of Academic Learner English* (CALE; Callies & Zaytseva 2013), the Michigan Corpus of Upper-level Student Papers (MICUSP; Römer and O'Donnell, B.M.) and the British Academic Written English (BAWE; Alsop & Nesi 2009). Preliminary findings reveal genre-induced differences in terms of the frequency of *although*-clauses in the writing of both groups and indicate variation in sentence positioning in research papers affected by writers' native language.

References

Ädel, A. 2008. Involvement features in writing: do time and interaction trump register awareness? In G. Gilquin, S. Papp & M. B. Diez-Bedmar eds. *Linking up Contrastive and Learner Corpus Research*. Amsterdam, Atlanta: Rodopi.

Alsop, S. & H. Nesi. 2009. Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, 4(1): 71-83.

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

Callies, M. & E. Zaytseva. 2013. The corpus of academic learner English (CALE). A new resource for the assessment of writing proficiency in the academic register. *Dutch Journal of Applied Linguistics*, 2 (1): 126-32.

Crew, W. 1990. The illogic of logical connectives. *ELT Journal*, 44 (4): 316-25.

Gillet, A. 2014. *Using English for Academic Purposes*. http://www.uefap.com/.

Granger, S. & S. Petch-Tyson. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15: 17-27.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot. 2009. *The International Corpus of Learner English. Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Kerz, E. 2013. Concessive adverbial clauses in L2 academic writing. In S. Granger, G. Gilquin & F. Meunier eds. *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead. (Corpora and Language in Use. Proceedings Vol 1)*. Louvain-la-Neuve: Presses universitaires de Louvain.

Morley, J. 2014. *The Academic Phrasebank*. http://www.phrasebank.manchester.ac.uk/ being-critical/.

Paquot, M., H. Hasselgård & S. Oksefjell Ebeling. 2011. Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. *Paper Presented at the International Conference on Learner Corpus Research 2011, Sept 2011, Louvain-la-Neuve, Belgium.*

Römer, U. 2009. English in academia: does nativeness matter? *Anglistik: International Journal of English Studies* 20 (2): 89-100.

Römer, U. & M. B. O'Donnell. 2011. From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora* 6(2): 159-77.

Siepmann, D., J. D. Gallagher, M. Hannay & J. L. Mackenzie. 2008. *Writing in English: A Guide for Advanced Learners*. Tübingen: Narr.

Wagner, S. 2011. Concessives and contrastives in student writing: L1, L2 and genre differences. In J. Schmied ed. *Academic Writing in Europe: Empirical Perspectives*. Göttingen: Cuvillier.

Posters

Discourse patterns – pattern discourse: The transdisciplinary potential of PATTERNS as a concept

Beatrix Busse & Ruth Möhlig-Falke (University of Heidelberg)

The term *pattern* is ubiquitous in linguistics, traditionally, however, with descriptive rather than theoretical significance. Since de Saussure (and before) linguists have used *pattern* to refer to recurring form-to-function mappings on the various levels of linguistic expression, the term usually not being specially defined.

Pattern has gained importance with the advent and growing influence of usage-based theories and corpus-linguistic methodology. While usage-based cognitive linguistics emphasises that the domain-general human capacity of pattern-finding is crucial for the acquisition of language (e.g. Ellis & Frey 2009, Tomasello 2000, 2003), corpus linguistics focuses on the question of how linguistic patterns (lexical or grammatical) can be found with the help of computer technology. In cognitive linguistics as well as in construction grammar, *pattern* is substituted by *schema* or (schematicised) *construction*, which may be understood as the underlying mental abstraction of observable patterns of language use (e.g. Langacker 2000, Goldberg 1995, Croft 2001). In corpus linguistics, patterns are typically understood in terms of *collocations* or *colligations* (e.g. Sinclair 1991, Hunston & Francis 2000), i.e. frequent (or in a weaker sense, recurrent) co-occurrences of content words or content-and-function words in different contexts of language use found in digitised corpora. Frequency of co-occurrence "[that] is larger than expected on the basis of chance" (Gries 2008: 6) becomes the main criterion for selecting those patterns (collocations, colligations, phraseologisms) that are worthy of further investigation (e.g. Stubbs 2002, Hunston & Francis 2000, Hoey 2005, Gries 2008, Ebeling & Oksefjell Ebeling 2013). As important as these approaches are for corpus linguistics in general, their understanding of *patterns* is limited in that it reduces them to observable sequences of words. As one of few, Stubbs (2013), however, points out that the ability to *find* order (or patterns) in these sequences has to do with a certain theory speakers have in their minds. At the same time, he stresses the need to integrate language into a theory of social structure, to assess what linguistic and semiotic patterns (on all levels of language and text) mean in contexts, which opens up a number of transdisciplinary research questions.

Our poster presents a work-in-progress report on an interdisciplinary research project currently under development at the University of Heidelberg. With its basis in English corpus linguistics (Busse 2010), this project brings together participants from the natural and social sciences as well as the humanities with the aim of analysing the concept of PATTERN in a trans- and interdisciplinary perspective. Transdisciplinary discourse has shown that the concept of PATTERN is understood radically differently across the disciplines and that mutual profit may be gained from this "pattern discourse" in terms of methodology, theory, application and interpretation of findings, with one of the main questions evolving around the conceptualisation of patterns as either construed by a (human) observer, or as self-existent (constructivism vs. empiricism). As corpus linguists we might ask: In the light of transdisciplinary findings on patterns, what actually counts as a pattern in relation to such conventionally central linguistic parameters and concepts like place, time, frequency, context,

dynamic variability, change and stability? What is the difference between *pattern* vs. *structure* or *construction*, *pattern* vs. *sequence* and *order*? Which other phenomena than those that have so far been considered as patterns in linguistics are linguistically relevant? This paper will present recent theoretical findings of the Heidelberg interdisciplinary research group on patterns and show by example how these have influenced quantitative and qualitative investigations of linguistic construction of urban space in Brooklyn, New York.

References

Busse, B. 2010. Recent trends in historical stylistics. In D. McIntyre & B. Busse, eds. *Language and Style*. London: Palgrave Macmillan, 32-54.

Croft, W. 2001. *Radical Construction Grammar*. New York: Oxford University Press.

Ebeling, J. & S. Oksefjell Ebeling. 2013. *Patterns in Contrast*. Amsterdam: Benjamins.

Ellis, N. C. & E. Frey. 2009. The psycholinguistic reality of collocation and semantic prosody(2): affective priming. In R. Corrigan, E. A. Moravcsik, H. Ouali & K. Wheatley eds. *Formulaic Language*, vol. 2: *Acquisition, Loss, Psychological Reality, and Functional Explanations*. Amsterdam: Benjamins, 473-97.

Goldberg, A. 1995. *A Construction Grammar Approach to Argument Structure*. Chicago: Chicago UP.

Gries, S. Th. 2008. Phraseology and linguistic theory: a brief survey. In S. Granger & F. Meunier eds. *Phraseology. An Interdisciplinary Perspective*. Amsterdam: Benjamins, 3-25.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hunston, S. & G. Francis. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins.

Langacker, R. 2000. *Grammar and Conceptualization*. Berlin: Mouton de Gruyter.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. 2002. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7: 215-44.

Stubbs, M. 2013. Sequence and order. The neo-Firthian tradition of corpus semantics. In H. Hasselgård, J. Ebeling & S. Oksefjell Ebeling eds. *Corpus Perspectives on Patterns of Lexis*. Amsterdam: Benjamins, 13-34.

Tomasello, M. 2000. The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences* 4/4, 156-63.

Tomasello, M. 2003. *Constructing a Language. A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard UP.

'The rock when heated gets expanded': Semantic and pragmatic innovations in ESL varieties

Eduardo Coto Villalibre (University of Santiago de Compostela)

Authors such as Chappell (1980: 411), Collins (1996: 43), Carter & McCarthy (1999: 54) and Anderwald (2014) describe the English *get*-passive as a linguistic puzzle, as a contentious point of discussion, and as the subject of widespread disagreement. This is because the various constructions containing *get* form a fuzzy set. In previous corpus-based research (Coto Villalibre 2014), I have tried to solve this fuzziness by placing these constructions on a gradient according to their degree of 'passiveness'. This gradient ranges from more to less passive and includes (i) central *get*-passives, in which the verbal past participle is dynamic, there can be an optional agent *by*-phrase, and an active equivalent is possible (*She got*

arrested by the police); (ii) pseudo *get*-constructions, with neither an active counterpart nor an agent phrase and whose past participle is stative (*They got married two days ago*); (iii) adjectival *get*-constructions, which show adjectival properties, e.g. premodification by a degree adverb (*I get very frustrated after a while*); (iv) idiomatic *get*-constructions (*get fed up with, get rid, of, get used to*); and (v) resultative *get*-constructions, with an intervening NP between *get* and the past participle (*He got himself shot*).

This paper discusses the formal, semantic and pragmatic characteristics of the various *get* constructions in three emerging varieties of English, namely Indian, Hong Kong and Singapore English; standard British English will also be analysed as the referent variety. The aim of the paper is, first, to confirm that *get* constructions are more common in the new varieties of English than in the native British variety, and, second, to throw some light on the reasons behind their frequent use through a study of their function. The corpora selected for the analysis are the corresponding spoken components of the *International Corpus of English* (ICE), which provide suitable material for contrastive studies of varieties of English worldwide. While the search for the different forms of *get* followed by a past participle was carried out automatically, further filtration was conducted manually. The inter-varietal examination will highlight relevant differences regarding the status of subjects (a shift in the animacy and responsibility features) and the semantics and pragmatics of the constructions as a whole. These differences point towards a tendency for *get* + past participle constructions to expand in English as a Second Language (ESL) varieties of English not only in frequency but also in function, since they can be seen to occur more freely in new environments and serve new functions. The influence of the substrate languages will also be taken into account in the analysis, since it seems to play a major role in the variation observed, especially with regard to the frequency of central *get*-passives in Indian English and to the higher incidence of *get* constructions with adversative meaning in Hong Kong English and Singapore English.

References

Anderwald, L. 2014. Getting acquainted, married, dressed, and shaved: passives or not? Paper presented at the *Third Conference of the International Society for the Linguistics of English (ISLE3)*, 24-27 August 2014, University of Zurich.

Coto Villalibre, E. 2014. From prototypical to peripheral: the '*get* + V*en*' construction in contemporary spoken British English. In A. Alcaraz-Sintes & S. Valera-Hernández eds. *Diachrony and Synchrony in English Corpus Linguistics*. Bern: Peter Lang, 205-31.

Carter, R. & M. McCarthy. 1999. The English *get*-passive in spoken discourse: description and implications for an interpersonal grammar. *English Language and Linguistics*, 3/1: 41-58.

Chappell, H. 1980. Is the *get*-passive adversative? *Papers in Linguistics: International Journal of Human Communication* 13/3: 411-52.

Collins, P. C. 1996. *Get*-passives in English. *World Englishes*, 15/1: 43-56.

Foreign words in interviews with EFL learners: Bridging lexical gaps?

Sylvie De Cock (Centre for English Corpus Linguistics, Université catholique de Louvain)

The Louvain International Database of Spoken English Interlanguage (LINDSEI) contains informal interviews with intermediate to advanced level learners of English as a foreign language. The interviews follow the same set pattern and are made up of three main tasks: a personal narrative based on a set topic (an experience that taught them a lesson, a country that impressed them, or a film or play they liked/disliked), a free discussion mainly about university life, hobbies, foreign travel or plans for the future and a picture description. Although the interviews are all conducted in English, "foreign" words sometimes feature in both the interviewers' and the learner interviewees' contributions.

This poster sets out to examine and compare the use of foreign words in five of the subcorpora included on the LINDSEI CD-ROM (Gilquin et al. 2010): interviews with Dutch, French-, German-, Italian- and Spanish-speaking EFL learners (LINDSEI_Dutch, LINDSEI_French, LINDSEI_German, LINDSEI_Italian, LINDSEI_Spanish). These subcorpora contain between 140,000 and 80,000 words of interviewee and interviewer speech. Foreign words have been specially marked up in the LINDSEI corpus (<foreign> WORD(S) </foreign). They were retrieved from the subcorpora using WordSmith Tools and carefully examined in context.

The study investigates the extent to which foreign words are used in the five learner varieties, whether or not the majority of the learner interviewees in each subcorpus under study resort to these items, the origin of the foreign words used (do the foreign words systematically come from the learners' mother tongue or do they sometimes come from another (foreign/second) language?) and the type of gaps the foreign words help learners bridge across the three interview tasks. The analysis reveals that, besides helping learners bridge vocabulary/lexical gaps (words/expressions that appear to be unknown or inaccessible to them; e.g. *cotizar*, *des algues*, *lasser*), the foreign words in LINDSEI very often serve as cultural/institutional bridges (e.g. *Tour de France*, *Parco Nazionale del Gran Paradiso*, *Vlaamse Opera*, *Abitur*, *gilles de Binche*) and pragmatic/discourse bridges (e.g. *ja*, *allez*, *si*, *enfin*, *bueno*).

Although the main focus of the study is on the foreign words used by the EFL learners, the analysis also explores the use of these words by the interviewers (most of whom are native speakers of English in the subcorpora under study) in the interactions: whether or not the interviewers initiate the use of some foreign words, how they react to the foreign words used by the learners, whether or not they repeat learners' foreign words and whether or not they provide the learners with any (lexical) help.

References

Gilquin, G., De Cock, S. & Granger, S. eds. 2010. *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

These points stated, a number of problems remain: A corpus-based analysis on the idiomatisation of collective noun-based constructions

Yolanda Fernández Pena (University of Vigo)

Collective nouns in English can collocate with singular or plural verbal forms depending on the speaker's focus, either on the collectivity or on the individual members, respectively (Levin 2001, 2006; Depraetere 2003; Hundt 2006, 2009). This alternation of verbal agreement is complicated even further if the collective takes a plural *of*-dependent, as in (1), which may interfere in the relationship of agreement between the collective controller and the verbal target (i.e. 'attraction'; Bock et al. 2001; Acuña-Fariña 2012):

(1) A random bunch$_{SG}$ of people$_{PL}$ are$_{PL}$ waiting […] [COCA: FIC Mov:Bean]

Plural *of*-dependents interact with verbal agreement to the extent that they very frequently determine the number of the verb and thus, as this study shows, they may even favour the progressive loss of the lexical status and content of the collective noun, thus leading to 'idiomatisation' (Akimoto 2002). In fact, this plural constituent has been attested as one of the factors involved in the grammaticalisation of constructions such as *a lot of* or *a bunch of* (Traugott 2008a, 2008b; Brems 2011; Traugott and Trousdale 2013). In this regard, this study contributes to the previous investigations on the field by exploring the influence of *of*-dependency in the collocational restrictions of homologous constructions such as *a number of* N_{PL}, *a group of* N_{PL} and *a/the majority of* N_{PL}, three structures which, in my previous studies on collective noun-based constructions, showed a remarkable frequency with plural dependents in standard English.

This paper broadens the scope of prior investigations by providing both a synchronic and a diachronic study of three collective noun-based constructions which covers both standard English and a selection of other inner-circle varieties. More specifically, the data were retrieved from the *British National Corpus* (BNC), the *Corpus of Contemporary American English* (COCA) and the different components of the *International Corpus of English* (ICE) for Canadian, Irish and New Zealand English. Besides, historical data attesting the evolution and lexical changes of these constructions were obtained from the *Corpus of Historical American English* (COHA).

The object of study here concerns two main issues: (i) the syntactic fixation of the construction and (ii) its (potential) grammatical meaning. In particular, the parameters that have been taken into account as possible indicators of the idiomatisation of the structures under scrutiny here pertain to their collocational restrictions: acceptability of premodification, specialisation with a specific determiner and the predominance of plural verbal agreement. Likewise, the instances retrieved from the historical corpus were individually analysed and contrasted with the contemporary sources so as to consider significant lexical changes across the span of time these corpora cover (19^{th}-21^{th} centuries).

The results obtained corroborate the tendencies observed so far as they suggest a clear syntactic fixation of the constructions explored here in both standard and regional English.

Thus, structures like *a number of* N_{PL}, *a majority of* N_{PL} and *a group of* N_{PL} take a remarkable rate of plural verbal forms (over 99%, 80% and 60%, respectively), a finding which evinces the loss of the lexical status of the collective noun in favour of the new controller of agreement, that is, the plural noun in the *of*-dependent. However, this drift towards the idiomatisation of the collective, already established in the 19th century, is only reinforced in the case of *a number of* by the quantificational meaning this element denotes and which points to the nearly idiomatic status of this expression in Present-Day English. *Group* and *majority*, by contrast, do not show such a clear grammatical nuance. Still, the trends discerned in this study evince the significance of *of*-dependency for the patterns of verbal agreement, an important observation which deserves further consideration in light of the already grammaticalised homologous constructions (i.e. *a lot of*).

References

Acuña-Fariña, J. Carlos. 2012. Agreement, attraction and architectural opportunism. *Journal of Linguistics*, 48(2): 257-95.

Akimoto, M. 2002. Two types of passivization of 'V+NP+P' constructions in relation to idiomatization. In T. Fanego, M. José López-Couso & J. Pérez-Guerra eds. *English Historical Syntax and Morphology: Selected papers from the eleventh International Conference on English Historical Linguistics (ICEHL 11), Santiago de Compostela, 7–11 September 2000*; vol 1: 9-22.

Bock, K., K. M. Eberhard, J. Cooper Cutting, A. S. Meyer & H. Schriefers. 2001. Some attractions of verb agreement. *Cognitive Psychology*, 43: 83-128.

Brems, L. 2011. *Layering of Size and Type Noun Constructions in English.* Berlin: Mouton de Gruyter.

Depraetere, I. 2003. On verbal concord with collective nouns in British English. *English Language and Linguistics*, 7(1), 85-127.

Hundt, M. 2006. The committee has/have decided…: On concord patterns with collective nouns in inner- and outer-varieties of English. *Journal of English Linguistics*, 34(3). 206-32.

Hundt, M. 2009. Concord with collective nouns in Australian and New Zealand English. In P. Peters, P. Collins & A. Smith eds. *Comparative Studies in Australian and New Zealand English: Grammar and Beyond*. Amsterdam: Benjamins, 207-24.

Levin, M. 2001. *Agreement with Collective Nouns in English*. Lund: Lund Studies in English.

Levin, M. 2006. Collective nouns and language change. *English Language and Linguistics*, 10(2): 321-43.

Traugott, E. C. 2008a. Grammaticalization, constructions and the incremental development of language: suggestions from the development of degree modifiers in English. In R. Eckardt, G. Jäger & T. Veenstra eds. *Variation, Selection, Development – Probing the Evolutionary Model of Language Change*. Berlin: Mouton de Gruyter, 219-50.

Traugott, E. C. 2008b. The grammaticalization of *NP of NP* patterns. In A. Bergs & G. Diewald eds. *Constructions and Language Change*. Berlin: Mouton de Gruyter, 23-45.

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

Measuring syntactic development in longitudinal learner corpora from a usage-based perspective

Kristopher Kyle & Scott Crossley (Georgia State University)

The linguistic features of L2 writing development and quality have been oft explored with regard to lexical sophistication (e.g., Kyle & Crossley, in press) and syntactic complexity at the clausal (e.g., Ortega, 2003; Lu, 2011) and phrasal (e.g., Biber, Gray, & Poonpon, 2011) levels. What has not yet been explored, despite the generally held notion in applied linguistics of the inseparability of lexis and grammar (e.g., Biber, Conrad, and Reppen, 1998) and the rising influence of usage-based perspectives to language acquisition (e.g., Ellis, 2002; Ellis & Ferreira-Junior, 2009), is the relationship between writing development and verb argument construction (VAC) use.

Usage-based theories of language posit that a form/meaning divide does not exist and that grammatical forms carry meaning in the same way as lexical items (e.g., Langacker, 1987; Goldberg, 1995). For example, in the sentence *He wugged her the ball*, we can extrapolate the meaning of the nonsense verb "wugged" because the ditransitive SVOO form carries the meaning of transferring something from one entity to another. Goldberg (1995) refers to these form-meaning pairings as *constructions*. Constructions occur at multiple levels of abstraction, ranging from morphology to syntax. VACs, or constructions that consist of a main verb and all arguments they take, have recently been of particular interest in L1 and L2 development research This research has suggested that input frequency and prototypicality are key indicators of how easily a particular VAC will be learned (e.g., Goldberg, Casenhiser, & Sethuraman, 2004; Ellis and Ferreira-Junior, 2009). VACs that are highly frequent in learner input tend to be learned more earlier, suggesting that high-frequency/prototypical VACs can be considered less sophisticated than low-frequency/prototypical VACs, much like current theories of lexical sophistication (e.g., Nation, 2006).

Most of the extant body of VAC development research explores a small number of constructions in relatively early stages of language development, leading to large gaps in our knowledge of linguistic development for all but the most salient constructions. Very little is known, for example, about the relationship between writing development and VAC use in either the L1 or the L2, despite concurrent interests in both written language development at the clausal level (e.g., Ortega, 2003) and usage-based language acquisition (e.g., Ellis, 2002). There are at least two important issues that help explain this gap, namely the lack of a comprehensive accounting of the constructions found in English (or any other language), and the absence of analysis tools. The first issue has recently begun to be addressed through the use of advanced natural language processing (NLP) techniques to identify and document the frequency profiles (based on the BNC) of VACs identified by Francis, Hunston, and Manning (1996) (e.g., Römer, O'Donnell, & Ellis, in press), and through the bottom-up approach I have used to create a comprehensive frequency database of the VACs in the 450 million-word Corpus of Contemporary American English (COCA). The second issue has also begun to be addressed through the development of the Tool for the Automatic Analysis of Syntactic

Complexity (TAASC), which can measure VAC frequency and prototypicality in learner texts.

These new developments allow for the current study, which investigates VAC development across a longitudinal corpus of free-writes written by six L2 English learners over a 12-month period studying at an intensive English program in the US. This study tests the hypothesis that learners will produce less frequent/prototypical VACs as their language skills develop. This has important implications for usage-based perspectives on language learning specifically and for second language acquisition in general.

CLiC Dickens: A cognitive corpus stylistic approach to patterns of body language presentation in fiction

Michaela Mahlberg, Peter Stockwell & Johan de Joode (University of Nottingham)

This poster will illustrate how the innovative online tool CLiC 1.0 (http://clic.nottingham.ac.uk/) can support the study of textual techniques of characterisation. The development of CLiC is set within a theoretical context that proposes a novel approach to the study of characterisation in Dickens by employing corpus linguistic methods within the framework of cognitive poetics. Cognitive poetics (Stockwell 2002) accounts for how readers take textual and inferential cues related to characters' behaviours and build up fictional minds surrounding the characters. The study of what is the crucial 'texture' (Stockwell 2009) of characterisation has not received sufficient attention to date. Our combination of corpus linguistics and cognitive poetics makes it possible to take a fresh view on language and characterisation in Dickens's novels. In corpus studies in general, prose fiction is typically treated as a single register without distinguishing between fictional speech and narration. With the help of corpus annotation, our approach makes patterns of character speech and descriptions provided by the narrator available for detailed study and comparison. In this poster we focus in particular on patterns of body language presentation. We use the concept of the suspension to focus on places in the text that are likely to contain body language. The concept of the 'suspension' (a stretch of narrator text that interrupts the speech of characters) is based on Lambert's (1981) 'suspended quotation'. Korte (1997) argues that presenting body language in contexts of speech, such as reporting clauses, can contribute to creating the naturalistic effect of simultaneity. Mahlberg & Smith (2012) and Mahlberg et al. (2013) have discussed some of the functions of patterns of suspensions. The present study extends the evidential basis for observations on body language. With the help of CLiC, we will present a systematic account of functions of body language specifically in Dickens – but also with reference to more general patterns in 19th century fiction. Among the patterns we discuss are examples of body language realised by practical actions (e.g. 'throwing himself back in his chair') as well as patterns where the emphasis is on the narrator's interpretation of the manner of speaking, as in (1) below (the suspension is highlighted in italics):

(1) ‘I know you would not mind,’ *said Agnes, coming to me, and speaking in a low voice, so full of sweet and hopeful consideration that I hear it now,* ‘the duties of a secretary.’ (Charles Dickens, *David Copperfield*)

In addition to presenting our interpretative account, the poster illustrates some of the functionalities of CLiC that we are developing to retrieve relevant data.

Imperatives vs. insubordinate *if*-clauses: A corpus study of directives in formal and informal spoken American English

Beatriz Mato-Míguez (University of Santiago de Compostela)

Directive meaning in English typically takes the form of an imperative clause, as in (1), or an imperative clause with the verb *let* followed by a subject in the objective case in situations in which the speaker includes himself/herself in the action proposed, as in (2). Conditional clauses can also code directives in uses in which the conditional meaning is marginal, if at all present, as shown in (3).

(1) Stay to your right folks, please, tour group coming out (*SBC* 040)

(2) Let’s talk about race...in terms of power. (SBC 011)

(3) If you could get them to me, I would be deeply appreciative (SBC 036)

From conditional clauses of the type in (3), a new construction seems to have arisen and become specialized in directive function, so-called insubordinate *if*-clauses, as illustrated in (4).

(4) Okay folks, if you will please take a look at this picture taken during construction (SBC 040)

Insubordinate *if*-clauses have undergone the process known as *insubordination*, that is, “the conventionalized main clause use of what, on *prima facie* grounds, appear to be formally subordinate clauses” (Evans 2007: 367; see also Evans & Watanabe forthcoming). Formally they are marked by the absence of a main clause which is not retrievable from the structural or situational context, their grammatical status thus being that of independent clauses. Prior research (Stirling 1999; Mato-Míguez 2014) has suggested that these clauses in English seem to be typically used to issue polite requests, since they imply an option with alternatives, the hearer being given the option of not complying with the action proposed. In my presentation, I will test this hypothesis on the basis of evidence extracted from the *Santa Barbara Corpus of Spoken American English* (*SBC*, 2000-2005) and the *Corpus of Spoken, Professional American-English* (*CSPAE*, 2000). I will analyze: (i) whether the use of insubordinate *if*-clauses is limited to certain directive categories or whether they can be used for all the directive meanings associated with imperative clauses (i.e. orders, offers, etc.) following for this purpose the classification of directive categories proposed by the standard reference grammars of English and by Pérez Hernández & Ruiz de Mendoza (2002); and (ii) whether

differences can be attested in the uses of imperatives, *let*-clauses and insubordinate *if*-clauses as regards the level of formality of the conversational exchange and the degree of politeness that the social distance between the speakers requires, this in order to address, among others, the following questions: which is the illocutionary force most frequently coded by each type of clause?; do imperative clauses and insubordinate *if*-clauses constitute true variant patterns, or have their uses become specialized to code specific directive meanings?; are insubordinate *if*-clauses more frequently employed for directive categories that go against the face of the addressee (i.e. orders, requests) than for illocutionary acts that are beneficial to the hearer (e.g. offers or invitations)?; are insubordinate *if*-clauses more frequently used in the *CSPAE* than in the *SBC* given the more polite nature of the type of interactions recorded in the former corpus?; are there conversational exchanges in which the three constructions co-occur and, if so, which is their sequential ordering?; are insubordinate *if*-clauses issued first to prevent a possible face-threatening act?

References

Barlow, M. 2000. *Corpus of Spoken, Professional American-English*. Rice University.

Du Bois, J. W., W. L. Chafe, C. Meyer, S. A. Thompson, R. Englebretson & N. Martey. 2000-2005. *Santa Barbara Corpus of Spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.

Evans, N. 2007. Insubordination and its uses. In I. Nikolaeva, ed. *Finiteness: Theoretical and Empirical Foundations*. New York: OUP, 366-431.

Evans, N: & H: Watanabe, eds. Forthcoming. *The Dynamics of Insubordination.* Amsterdam: Benjamins.

Mato-Míguez, B. 2014. *If you would like to lead*: On the grammatical status of directive isolated *if*-clauses in spoken British English. In A. Alcaraz-Sintes & S. Valera-Hernández, eds. *Synchrony and Diachrony in English Corpus Studies*, 259-83. Bern: Peter Lang.

Pérez H., L. & F. Javier Ruiz de Mendoza. 2002. Grounding, semantic motivation, and conceptual interaction in indirect speech acts. *Journal of Pragmatics* 34: 259-84.

Stirling, L. 1999. Isolated *if*-clauses in Australian English. In P. Collins & D. A. Lee, eds. *The Clause in English.* Amsterdam: Benjamins, 273-94.

Translationese or transfer: Disentangling translation effects from L2 learning effects

Stella Neumann, Elma Kerz & Marcus Ströbel (RWTH Aachen University)

Baker's (1993) suggestion that translations need to be investigated in their own right, i.e. separately from original texts, triggered a host of corpus research to identify translation properties (cf., e.g., papers in Mauranen and Kujamäki 2004). Typically, these first generation studies were limited in scope in terms of e.g. languages involved, features investigated, corpus size or sometimes also in definition and operationalisation of translation properties. In recent years, therefore, the claim that certain features are specific to translations has been called into question (e.g. Becher 2010 on explicitation). At the same time, computational approaches have been successful at spotting so-called translationese, i.e. linguistic patterns so typical of translations that machine learning techniques can reliably discriminate translations from originals (e.g. Baroni and Bernardini 2006; Volansky et al.

2015) and at discovering latent dimensions separating translated from non-translated texts drawing on a suite of multivariate techniques (Diwersy et al. 2014). While these findings are impressive, they do not account for the relationship to other modes of multilingual language use that might prove similar to prototypical translation (Halverson 2013) including learner language but ultimately also L2 varieties (cf. Mukherjee and Hundt 2011).

In their discussion of the CroCo Corpus of English and German originals and translations, Hansen-Schirra and Steiner (2012) link translation to other forms of language contact and point out that, depending on whether a text is translated into the L1 or the L2, it may involve borrowing or shift phenomena respectively suggesting that attempts to understand translation phenomena should consider the direction of translation: The less common case of translating into the L2 could give rise to different phenomena than more common cases of translating into the L1. Patterns observed in other types of L2 writing, e.g. in contact varieties or academic registers, could be assumed to be different from the default case. Furthermore, the close relationship of a translated text to its source text appears to set apart translation from other types of multilingual language use attested, for example, in journalistic writing and other instances of recoding information in a text for a different audience within one and the same language.

The present paper sets out to test claims about translation-inherent properties empirically drawing on a corpus combining originals and translations in the language pair English and German with advanced L2 productions of German learners of English, thus affording the systematic investigation of patterns specific to one of the two modes as well as more general patterns applying to both. To this end, both translations proper and L2 productions are described relative to various alleged translation-specific indicators and general indicators of language proficiency proposed in the second language learning literature with the primary focus on a number of measures of lexical and syntactic complexity (cf., e.g., Connor-Linton and Polio 2014; Housen et al. 2012; Ortega 2003) and then analysed using techniques from machine learning. The proposed methodology permits inferences about the extent to which translationese amounts to general L2 transfer (or interference) effects.

References

Baker, M. 1993. Corpus linguistics and translation studies. Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli eds. *Text and Technology. In Honour of John Sinclair.* Amsterdam: Benjamins, 233-50.

Baroni, M. & S. Bernardini. 2006. A new approach to the study of translationese: machine-learning the difference between original and translated text. *Literary and Linguistic Computing,* 21(3): 259-74. doi:10.1093/llc/fqi039.

Becher, V. 2010. Abandoning the notion of translation-inherent explicitation: against a dogma of translation studies. *Across Languages and Cultures*, 11 (1): 1-28. doi:10.1556/Acr.11.2010.1.1.

Connor-Linton, J. & C. Polio. 2014. Comparing perspectives on L2 writing: multiple analyses of a common corpus. Special Issue. *Journal of Second Language Writing,* 26: 1-106.

Diwersy, S., S. Evert & S. Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In B. Szmrecsanyi & B. Wälchli eds. *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech.* Berlin/New York: de Gruyter. 174-204.

Halverson, S. L. 2013. Implications of cognitive linguistics for translation studies. In A. Rojo & I. Ibarretxe-Antunano eds. *Cognitive Linguistics and Translation Advances in Some Theoretical Models and Applications*. Berlin, Boston: de Gruyter, 33-74

Hansen-Schirra, S. & E. Steiner. 2012. Towards a typology of translation properties. In S. Hansen-Schirra, S. Neumann & E. Steiner eds. *Cross-Linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German.* Berlin: de Gruyter Mouton, 255-79.

Housen, A., F. Kuiken & I. Vedder eds. 2012. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: Benjamins.

Mauranen, A. & P. Kujamäki eds. 2004. *Translation Universals. Do they Exist?* Amsterdam: Benjamins.

Mukherjee, J. & M. Hundt eds. 2011. *Exploring Second-Language Varieties of English and Learner Englishes. Bridging a Paradigm Gap*. Amsterdam: Benjamins.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing. *Applied Linguistics,* 24(4): 492-518.

Volansky, V., N. Ordan& S. Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 30(1): 98–118. doi:10.1093/llc/fqt031.

Language Change Database: A new online resource

Terttu Nevalainen, Tanja Säily & Turo Vartiainen (University of Helsinki)

A major challenge in linguistic research is unravelling the process of language change. Sociolinguists have made great strides in analysing change in apparent time, comparing the language use of successive generations at a given point in time. Information on real-time change is much sparser, but thanks to the digital turn in the humanities, more data is available that enables the diachronic approach. However, empirical work on change over time is still fragmented, and some of it is not easily available.

In our new funded project, we propose to make empirical linguistic research on language change more cumulative by compiling an online, open-access Language Change Database (LCD) to provide comparative, real-time baseline data on change in progress, starting with seminal studies that analyse both linguistic and register changes based on digital corpora. This annotated database, which is currently being compiled by the team members in international collaboration, will include research on the details of the diffusion of a variety of changes from Old English to the present day. Like the Corpus Resource Database (CoRD), it will have an easily searchable basic structure that allows rapid data retrieval across a number of parameters. Where available and copyright permitting, the LCD will include the original data tables in a downloadable format as well as links to the original papers and research reports. The database will be published on a wiki-style platform where scholars around the world will be able to collaboratively update it with their own work.

The justification for the Language Change Database is twofold. First, we argue that having access to an extensive and cumulative resource on past work on real-time language change will contribute to a better understanding of the ways in which languages change across communities, varieties, registers and language systems over time. Secondly, it will help bridge the gap between those who know the historical data first hand and those who are working on more abstract models using statistical methods, or need historical baseline data to

better understand present-day variation. This annotated resource will bring together in one repository a wealth of research that is currently scattered and fragmented, and often only available in print publications.

The database will be part of the e-resources of the VARIENG Research Unit, and we hope that it will serve as the basis for future work ranging from statistical modelling and systematic reviews to the replication of earlier research with other data sets. Our paper introduces the project and the solutions that have so far been adopted in the planning of the structure of the database. We welcome any ideas and suggestions from the conference participants concerning any pilot projects on the beta version of the database.

References

CoRD = Corpus Resource Database. http://www.helsinki.fi/varieng/CoRD/.
LCD = Language Change Database. http://www.helsinki.fi/lcd/.
VARIENG = Research Unit for the Study of Variation, Contacts and Change in English. http://www.helsinki.fi/varieng/.

Precious few and *practically all*: The modification of absolute and relative quantifiers

Ngum Meyuhnsi Njende (KU Leuven)
Kristin Davidse (KU Leuven)
Lobke Ghesquière (KU Leuven / Research Fund Flanders / Vrije Universiteit Brussel)

There is a long tradition of research into the modification of adjectives, e.g. *very pretty*, *almost full* (Paradis 2001). By contrast, little focused attention has gone to the modification of quantifiers, e.g. *very few*, *almost all*. As part of a larger project aimed at filling this gap, we present a corpus-based study of the modification of *(a) few* and *all*.

Quantification can be described in terms of two basic types: absolute and relative quantification (Milsark 1977, Langacker 1991). *Absolute* quantifiers *measure* the size of some set or mass with reference to an implied scale with (schematic) measure units, e.g. *(a) few*, *many/much*. *Relative* quantifiers *compare* the size of the actually designated mass or set to a reference mass/set, indicating whether or not the two coincide, and if not, to what extent they don't, e.g. *NO*, *SOME*, *most*, *all*.

Ghesquière and Davidse (2011) proposed that, just as the distinction between unbounded and bounded adjectives determines their different types of degree modification, viz. scalar (e.g. *very pretty*) versus proportional (e.g. *almost full*), the distinction between absolute and relative quantifiers motivates different, analogous types of quantity modification. Therefore, we hypothesize that Quirk et al.'s (1985) subclasses of degree modifiers can be reconceptualized into subclasses of quantity modifiers.

With ***absolute*** quantifiers, modifiers operate on the 'range' indicated on the scale, which they 'upscale' (boosters), e.g. *very many*; 'downscale' (diminishers), e.g. *so little*; or 'hedge'

(compromisers), e.g. *rather a lot* (see Quirk et al. 1985). This scaling effect interacts with the positive or negative scalar direction inherent in the quantifier. Thus, with positive *a few*, we predict that boosters will enlarge the quantity (1a). Conversely, with negative *few*, boosters will further reduce the quantity (2a) while compromisers will soften the paucity meaning of the quantifier (2b).

(1) I got offered *quite a few* bad guys in American movies (WB)

(2) a. Then look at the *precious few* victories (WB)

b. When they look at senior women in their organisations they cannot help but notice that *rather few* of them have rich family lives. (WB)

With ***relative*** quantifiers, modifiers operate on the (non-)coincidence with the reference mass indicated. With *all* they indicate that the quantity designated indeed coincides with the reference mass/set (maximizer), e.g. *absolutely all* (3a), or differs somewhat from it (approximator), e.g. *practically all* (3b), thus mitigating the universal quantification that *all* designates on its own.

(3) a. if *absolutely all* other options fail. (WB)

b. disabled secretaries can do *practically all* office jobs. (WB)

We also investigate the hitherto largely neglected possibility of a ***marked relative*** reading being imposed on ***absolute*** quantifiers so that the quantity they indicate is construed in relation to a reference mass (Milsark 1977), as in (4), where (stressed) *a FEW* is equivalent to relative *SOME*, indicating that the actual set differs considerably from the reference set of *all ten candidate countries*. We will investigate if the submodification of these marked uses differs in any way from that of the congruent absolute uses of *(a) few*.

(4) If *only a few*, but not all ten candidate countries, are ready by December (WB)

We will test the above hypotheses by qualitative and quantitative analysis of large datasets extracted from the Times subcorpus of Wordbanks*Online* (WB). For *(a) few* an exhaustive extraction of 19,965 was made, of which 1,968, i.e. roughly 10%, had quantity modification. For *all* a random sample of 20,000 was taken, of which only 208, about 1%, was modified. This general difference in relative frequency, and the different proportions of the submodifier classes, will be incorporated in our generalizations about the different types of modification found with absolute and relative quantifiers.

References

Ghesquière, L. & K. Davidse. 2011. The development of intensification scales in noun-intensifying uses of adjectives: sources, paths and mechanisms of change. *English Language and Linguistics*, 15 (2): 251-77.

Langacker, R. W. 1991. *Foundations of Cognitive Grammar*, vol 2: *Descriptive Application*. Stanford: Stanford University Press.

Milsark, G. 1977. Toward an explanation of certain peculiarities of the existential construction in English. *Linguistic Analysis*, 3: 1-30.

Paradis, C. 2001. Adjectives and boundedness. *Cognitive Linguistics*, 12: 47-65.

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

How can you feel better than terrific? Insights on the history of *terrific*

Paloma Núñez-Pertejo (University of Santiago de Compostela)

In line with the development undergone by a number of adjectives in the history of English, such as *awesome*, *bare*, *brutal*, *massive* and *wicked*, *terrific* has come to express positive meanings when it originally conveyed negative ones (cf. Mair 2006: 38-39; Robinson 2010). This lexical semantic change, which seems to be common with words denoting an evaluative meaning, has been referred to in the literature as '(a)melioration', 'elevation' or 'improvement of meaning' (cf. Culpeper 1997; Schendl 2001; Trask 2007). The tendency to endow originally negative words with positive meanings is particularly noticeable in the language of teenagers (Rodríguez González & Stenström 2011; Palacios Martínez & Núñez Pertejo 2012), but can also be observed in English generally.

The aim of this paper is to describe the syntactic and semantic variation of the polysemous adjective *terrific* based on the evidence afforded by *The Corpus of Late Modern English Texts* (CLMET3.0; cf. De Smet, Diller & Tyrkkö 2013), with a special emphasis on the frequency, function(s), meaning(s) and main collocations of this adjective.

A preliminary analysis of the data shows that three main basic meanings can be identified for *terrific*: (i) 'terrible, frightful'; (ii) 'of great size', 'tremendous'; (iii) 'excellent, amazing'. However, and as very often happens in processes of semantic change in general, these senses may at times overlap, so that not all examples can safely be ascribed to one of the above groups, and thus need to be analysed independently. As regards frequency and development, my results so far show that *terrific*, a loanword from French first attested in 1667 according to the OED, remained quite uncommon for a long time (only 4 tokens are recorded in subperiod 1, covering 1710-1780, of CLMET3.0), and was used mostly in its negative meaning of 'terrible, frightful' (e.g. *called out aloud in a most* ***terrific*** *voice*). During subperiod 2 (1780-1850), *terrific* greatly increases its frequency (108 tokens), and from the second half of the 19th century onwards, starts to be used more often in sense (ii) above ('of great size', 'tremendous'; e.g. *at a* ***terrific*** *pace*). The first examples of *terrific* with a positive meaning ('amazing' 'excellent'; e.g. *the reaction was genuinely* ***terrific***) are not attested until the 19th century, and we have to wait until well into the 20th century for this meaning to become widespread, giving rise to what Mair calls "an all-purpose positive evaluator" (2006: 47). A further step in the developmental cline of *terrific* is represented by its increasing ability to occur on its own, as an emotive manifestation of approval (e.g. *'Thanks awfully,' said Rex. 'That'll be ripping.' 'Fine!' said Derek Yardley. 'Great!* ***Terrific****!'*).

References

Culpeper, J. 1997. *History of English*. London & New York: Routledge.

De Smet, H., H.-J. Diller & J. Tyrkkö. 2013. *The Corpus of Late Modern English Texts*, version 3.0 (CLMET3.0). More information: https://perswww.kuleuven.be/~u0044428/clmet3_0.htm.
Mair, C. 2006. *Twentieth Century English*. Cambridge: Cambridge University Press.
Palacios Martínez, I. M. & P. Núñez Pertejo. 2012. *He's absolutely massive. It's a super day. Madonna, she is a wicked singer*. Youth language and intensification: A corpus-based study . *Text & Talk*, 32(6): 773-96.
Robinson, J. 2010. *Awesome* insights into semantic variation. D. Geeraerts, G. Kristiansen & Y. Peirsman eds. *Advances in Cognitive Sociolinguistics*. Berlin/New York: De Gruyter, 85-109.
Rodríguez González, F. & A.-B. Stenström. 2011. Expressive devices in the language of English- and Spanish-speaking youth. *Revista Alicantina de Estudios Ingleses*, 24: 235-56.
Schendl, H. 2001. *Historical Linguistics*. Oxford: Oxford University Press.
Trask, R. L. 2007. *Historical Linguistics*. 2nd ed. London: Arnold.

Law, administration, language and translation: Creating a multilingual corpus

Arja Nurmi, Marja Kivilehto, Annikki Liimatainen & Anna Ruusila (University of Tampere)

The study of language for special purposes and the study of translation for special purposes is lacking in language-specific research. Our research project attempts to correct this by looking at Finnish texts in comparison to corresponding English, German and Swedish ones, and, to a minor extent, also French and Spanish. In order to carry out our research goals, we are planning a multilingual corpus. Our main focus is legal and administrative texts, with individual scholars working on specific language pairs, with a varying focus on either law or both law and administration.

The corpus is centred around a hub of Finnish texts, both legal and administrative. It is complemented by corresponding texts in the other languages we study. The corpus is intended to be both parallel and comparable, i.e. to contain both translations to and from Finnish and texts originally written in all the languages studied within the project. This will allow the comparative study of writing practices in different languages for different audiences as well as the study of translation. The results of our studies will help translators in solving actual problems in their work, making domain-specific text conventions visible, and also allow for more research-based teaching of legal and administrative translation.

The research questions at the heart of our project include features of language and text such as conventionalised phrases (e.g. routine formulae, collocations, word pairs and archaisms), formulaic texts (e.g. oaths), modality, reader–writer relationship, sentence structure and terms and terminology. This is by no means an exhaustive list. We will apply both qualitative and quantitative methods in our studies, and draw from many fields of research, including translation for specific purposes, LSP research, phraseological research, contrastive linguistics, legal linguistics, text analysis, terminology and audience design.

The poster will mainly illustrate the planned structure of the corpus with regard to Finnish and English. Many of the texts to be included are already available online in electronic form,

and the purpose of building a corpus is to allow for a systematic study, using proper corpus tools.

Verbs, verbs, verbs: The progressive vs. non-progressive paradigm in World Englishes

Paula Rautionaho (University of Tampere)

This poster reports on the results of a recent study into the progressive in World Englishes (Rautionaho 2014), and argues in favour of re-considering the measuring of the frequency of the progressive. In most previous studies, the number of progressives has been normalised to 100,000 *words* (so-called M-coefficient). However, a more accurate picture is gained by considering the number of progressives against the number of *verbs* (so-called V-coefficient, see Smitterberg 2005). The V-coefficient takes into account the fact that progressives can only occur in a verb phrase, not in a noun phrase or an adverbial phrase, and is thus able to give a more accurate measurement of the frequency of the progressive. Rautionaho (2014) uses the M-coefficient for comparing the frequency of progressives to results obtained in previous studies, and the V-coefficient for comparing the number of progressives in the World Englishes included in the study (BrE, IrE, AmE, JamE, IndE, PhiE, SinE and HKE).

In addition to providing more accurate frequency measures, considering the progressive vs. non-progressive paradigm offers a mass of information on the verbal paradigm in more general terms. For instance, Rautionaho (2014) shows that some of the varieties investigated prefer the present tense with the progressive (e.g. IndE), while for others, the proportion of present tense forms is higher with the non-progressive (e.g. IrE). This may give an indication of the integration of the progressive into the varieties in question. Moreover, the investigation of the progressive vs. non-progressive paradigm with regard to the use of modal auxiliaries indicates that there are differences in the use of the modal auxiliaries with the progressive: non-progressive modal verb phrases are clearly more evenly distributed in the varieties under investigation, with proportions ranging from approximately 10% to 14%, while the proportions of corresponding progressive verb phrases are more varied, ranging from 2% to 12%. Again, this may be considered to relate to the integration of the progressive on the one hand, and on the other, to the general trends related to the use of modal auxiliaries in World Englishes.

The poster thus argues for the re-consideration of measuring the frequency of the progressive: it is more accurate and more fruitful to compare the number of progressives to the number of verbs rather than words in a corpus. The V-coefficient offers the researcher a more precise tool to work with, as well as providing information on the progressive vs. non-progressive paradigm on a more general level. The benefits of using the V-coefficient are illustrated with the help of a recent study into the progressive in World Englishes. With the help of data retrieved from the *International Corpus of English*, Rautionaho (2014) fills in two gaps in the already vast research on the progressive: it focuses on spoken data rather than written data, and investigates a large number of World Englishes instead of focusing on one or two

varieties. It thus provides the first large-scale comparison on the frequency of the progressive, its morphosyntactic and lexical environment and its semantic functions.

References

Rautionaho, P. 2014. *Variation in the Progressive: A Corpus-Based Study into World Englishes.* Tampere University Press: Acta Universitatis Tamperensis, 1997.

Smitterberg, E. 2005. *The Progressive in 19th-century English: A Process of Integration*. Amsterdam: Rodopi.

Variation in the use of light verbs in varieties of English

Patricia Ronan (Université de Lausanne)

In light verb constructions we typically find a semantically low content, high frequency verb, such as *have, make, give* or *take,* which is predicated by a, typically verb-derived, predicate noun such as *to make a proposal*, or *to have a chat.* These constructions are semantically non-compositional and can typically be replaced by a verbal simplex such as *to propose* or *to chat.* These constructions are known from a variety of languages such as German (von Polenz 1987), French (Danlos 1992) and indeed English (e.g. Allerton 2002). In the proposed paper we will investigate to what extent variation in the use of light verbs can be found in different varieties of contemporary English.

It has previously been observed that light verb constructions are by no means identical in all varieties of a language. For the English language, it has been noted that the use of the light verb *have* is increasingly more popular in British English than in American English, whereas the light verb *take* is rising in American English (Algeo 1995, Dixon 2005: 461, Leech, Hundt and Mair 2009: 176), e.g. in *to have a swim* and *to have a look* versus *to take a swim* and *to take a look*. It has further been found that *give* is used more frequently in British than American English (Leech, Hundt and Mair, ibid.), and that Asian varieties show considerable variation in light verb use (Hoffman, Hundt and Mukherjee 2011).

Variety-specific selection preferences for light verbs are likely to be determined by the semantic properties of the light verb in question (e.g. Allerton 2002: 176-209) and new constructions are likely to be added to existing ones on the basis of semantic similarities. Thus the existence of constructions like *take a walk* or *take a stroll* is likely to allow for the creative extension of the light verb *take* to new contexts of denoting movement like *take an amble*. This qualitative and quantitative study will show what variety specific differences can be observed in the use of light verbs in different varieties of contemporary English. This is done by comparing semantic fields of predicate nouns in light verb constructions that are found with select high-frequency light verbs.

Data are drawn from the subcorpora of the *International Corpus of English* family, which represent both L1 and L2 varieties of English. Attestations will be culled from the corpora by

way of searching for the relevant light verbs with the help of the concordance program AntConc (Anthony 2005).

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*, 7-13.

Algeo, J. 1995. Having a look at the expanded predicate. In: B. Aarts & Ch. F. Meyer eds. *The Verb in Contemporary English. Theory and Description*. Cambridge: Cambridge University Press, 203-17.

Allerton, D. J. 2002. *Stretched Verb Constructions in English*. London: Routledge.

Danlos, L. 1992. Support verb constructions: linguistic properties, representation, translation. *French Language Studies*, 2(1): 1-32.

Dixon, R. M. W. 2005. She gave him a look, they both had a laugh and then took a stroll. GIVE A VERB, HAVE A VERB and TAKE A VERB constructions. In R.M.W. Dixon ed. *A Semantic Approach to English Grammar*. Oxford: Oxford University Press. 459-83.

Hoffmann, S., M. Hundt and J. Mukherjee. 2011. Indian English – an Emerging Epicentre? A Pilot Study on Light-Verbs in Web-derived Corpora of South Asian Englishes. *Anglia*, 12(3-4): 258-80.

Leech, G., M. Hundt, Chr. Mair. 2009. *Change in Contemporary English*. Cambridge: Cambridge University Press.

von Polenz, P. 1987. Funktionsverben, Funktionsverbgefüge und Verwandtes. Vorschläge zur Satzsemantischen Lexikographie. *Zeitschrift für germanische Linguistik*, 15(2): 169-89.

Null subjects in spoken English: ICE-GB and ICE-Singapore in contrast

Verena Schröter (Albert-Ludwigs-Universität Freiburg)

The so-called "New Englishes" are one of the most fruitful areas of study in corpus-based linguistics. The wealth of authentic language data provided by the ICE family enables the empirical evaluation of long-standing hypotheses on language contact and its linguistic consequences. Their size allows for the investigation of low-frequency phenomena while the wide availability of the corpora facilitates exchange between researchers and makes their results accountable. This poster presents a comparative study of two distinct varieties of English based on the respective ICE-corpora with regards to a classic typological parameter, that of null subjects.

According to reference grammars and typological classifications, Standard English is a language with obligatory subject pronouns. Still, especially in informal spoken mode, rare occurrences of empty subject slots can be observed (e.g. Wagner 2012).

In contrast, null subjects are commonly described as a distinctive feature of Colloquial Singapore English ('Singlish'), typically explained by contact with local substrates (e.g. Ansaldo 2009, Bao 2010, Lim 2004), but so far there is no transparent systematic, data-based study of their extent and conditioning.

In this study, variable expression of subject pronouns in the spoken component of ICE-SIN is compared with the parallel subcorpus of ICE-GB as a "standard English" reference. The

systematic comparison attests radical differences. Non-canonical British null subjects occur rather sporadically (less than 2%); they are largely confined to utterance-initial position in main clauses, or to recurring expressions such as *don't know*, *can't remember*, etc. (confirming earlier findings by Torres Cacoullos & Travis 2014, Weir 2009).

Resisting all prescriptive interventions, Singlish null subjects are found in diverse lexical, phonological, and syntactic environments and thus form a viable, if minor, grammatical option (around 10%). Examples include null subjects in main clauses (1), in sub-clauses co-referential with the subject of the matrix clause (2), in sub-clauses with different reference than the matrix-clause subject (3), or in sub-clauses co-referential with the null subject of the matrix clause (4).

(1) $ø_{(i)}$ Probably save quite a lot if $you_{i)}$ open it up to any doctor. 005#44

(2) Charles Dickens I read Christmas Carol that was that was cute not too bad so $I_{(i)}$ think $ø_{(i)}$ can understand. 090#402

(3) Because $they_{(i)}$ say $ø_{(i/?)}$ cannot switch on air-con, so $ø_{(j)}$ have to sleep without air-con. 014#184

(4) $ø_{(i)}$ Don't know whether $ø_{(i)}$ got time. 032#187

Their distribution is conditioned by factors commonly described for Chinese null subjects such as discourse salience and information status of the referent (Li & Thompson 1981, Loar 2011). This is especially visible in the frequent clusters of null subjects in the Singlish conversations that resemble Chinese topic chains.

References

Ansaldo, U. 2009. The Asian typology of English: theoretical and methodological considerations. In L. Lim & N. Gisborne eds. *The Typology of Asian Englishes.* Special Issue of *World Englishes,* 26: 133-48.

Bao, Z. 2010. A usage-based approach to substratum transfer: the case of four unproductive features in Singapore English. *Language,* 86: 792-820.

Li, C. & S. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Lim, L. ed. 2004. *Singapore English*: *A Grammatical Description*. Amsterdam: Benjamins.

Loar, J. K. 2011. *Chinese Syntactic Grammar*: *Functional and Conceptual Principles*. New York: Peter Lang.

Torres Cacoullos, R. & C. E. Travis. 2014. Prosody, priming and particular constructions: the patterning of English first-person singular subject expression in conversation. *Journal of Pragmatics,* 63: 19-34.

Wagner, S. 2012. *Null Subjects in English. Variable Rules,Variable Language?* Postdoctoral Thesis, Chemnitz University of Technology.

Weir, A. 2009. Subject pronoun drop in informal English. [Manuscript.]

Detecting lexical patterns in an English-language corpus of suicide notes

Jess Shapero (Freelance English language consultant, UK)
Sue Blackwell (Language Consultancy Desk, UK)
Iman Laversuch (Universität zu Köln)

According to the World Health Organization (2014), each year over 800,000 people commit suicide, making it the second leading cause of death amongst adolescents and adults aged 15-29 worldwide. To help shed light upon the thoughts and emotions of suicidal individuals, social scientists have traditionally relied upon analysing notes left behind. Given the relative difficulty in obtaining access to this data, linguists interested in investigating suicide have traditionally utilized the few pre-existing collections of both genuine and simulated suicide notes which have been available to the general public (e.g. Shneidman & Farberow 1957). Although these corpora have been invaluable in understanding the suicide note as a forensic text type, the majority of these compilations were gathered from comparatively limited sets of authors and have since become relatively dated. Consequently, the patterns identified in these corpora may not be representative of the language norms of contemporary suicidal individuals or those from other language groups.

To address these limitations, the authors of this study are in the process of compiling an annotated international corpus of suicide notes. This corpus will, it is hoped, grow to become a multilingual one and eventually form part of a wider corpus of text-types relevant to research in Language and Law. For now, however, we are working with English-language suicide letters originating from New Zealand, the UK and USA.

Such a corpus could be used by linguists to identify commonalities and potentially statistically significant differences in the content, framework and format of suicide notes across different demographic variables (e.g. gender, age, nationality, ethnicity/race, profession, etc.). The findings garnered from these analyses would also be of immense potential value to a wide range of professionals including lawyers, psychologists, and police officers. For example, the findings could be used to facilitate comparisons between falsified and genuine suicide notes, and to provide valuable insights into the psychological states of suicidal individuals. However, the collection and storage of such sensitive data present a number of significant challenges.

This presentation will discuss the ethical and practical issues which have to be addressed in compiling a corpus of suicide notes. It will also describe the decisions taken regarding the transcription, encoding, and annotation of the notes along with the relative merits of various software packages (including WordSmith Tools for lexical analysis, and The General Inquirer and Wmatrix for semantic tagging) used to analyse them. Building on the work of Shapero (2011), we will attempt to establish whether it is possible to define the suicide note as a discrete text-type and whether such notes are closer to spoken or written discourse. These questions have become particularly important given the dramatic increases which have been observed internationally in the incidence of suicide notes posted on social media: email, Twitter, Facebook, etc. (e.g. Barak & Miron, 2005; Ruder et al. 2011; Tam et al. 2007). Furthermore, the comparison of multiple notes written by single authors as well as between

different authors will enable insights into the range of inter- and intra- subject linguistic variation. To investigate these issues, we will be examining the lexical density, collocations and semantic themes which have been identified as being relevant in the commission of suicide (e.g. Leenaars, 1992; Leenaars et al. 2003).

References

Barak, A. & O. Miron. 2005. Writing characteristics of suicidal people on the Internet. *Suicide and Life-Threatening Behavior*, 35: 507-24. doi: 10.1521/suli.2005.35.5.507.

Leenaars, A. 1992. Suicide from Canada and the United States. *Perceptual and Motor Skills*, 74, 278. doi: 10.2466/pms.1992.74.1.278.

Leenaars, A., J. Haines, S. Wenckstern, C. Williams & D. Lester. 2003. Suicide notes from Australia and the United States. *Perceptual and Motor Skills*, 92: 1281-82. doi: 10.2466/pms.2003. 96.3c.1281.

Ruder, T., G. Hatch, G. Ampanozi, M. Thali & N. Fischer. 2011. Suicide announcement on facebook. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, 32: 280-82. doi: 10.1027/0227-5910/a000086.

Shneidman, E. & N. Farberow. 1957. *Clues to Suicide*. New York: McGraw Hill.

Shapero, J. 2011. *The Language of Suicide Notes*. Doctoral thesis, University of Birmingham, Birmingham, England. http://etheses.bham.ac.uk/1525/1/Shapero11PhD.pdf.

Tam, J., W. Tang, & D. Fernando: 2007. The internet and suicide. *European Journal of Internal Medicine*, 18: 453-55. doi:10.1016/j.ejim.2007.04.009.

World Health Organization 2014. *Mental Health: Suicide Data*. http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/.

Variation in the use of *in* as a spatial preposition in Ugandan English

Jude Ssempuuma (Ruhr University Bochum)

This paper explores the use of *in* as a spatial preposition in Ugandan English and the extent to which it varies from Standard English. Schmied (2008: 456) writes that the use of the most frequent English preposition *in* at the expense of *into* "occurs more significantly in East African than in British English". Mwangi (2004: 27) remarks variations from Standard English in the use of *in* as a spatial preposition in Kenyan English is an influence from Kiswahili and Gikuyu languages which make no distinction between prepositions of position and destination. Preposition *in* is the most frequently used of the 23 prepositions analyzed in the Ugandan data, namely; *in, of, to, for, from, at, with, on, up, by, about, around, among, over, into, off, under, within, down, near, onto, inside,* and *round,* with the frequency of 19.67 per 1,000 words.

This paper shows that in Ugandan English the variation from Standard English in the use of preposition *in* mainly involves the use of *in* instead of *at* and *to* other than *into* as claimed by Schmied quoted above. In addition, unlike in Kenyan English where Kiswahili and Kikuyu seem to be the substrate languages to influence variation from Standard English in the use of preposition *in,* in Ugandan English it seems that the influence is from the substrate indigenous Ugandan languages such as Luganda, Runyakole-Rukiga and Luo. For instance, Mufwene (2013: 218) observes that "in former exploitation colonies, we can still identify the

sources of the relevant substrate influences, how they compete with each other, and what are the external and internal ecological factors that bear on the selection process".

The Ugandan spoken data used for the analysis was compiled using a balanced or sample corpus approach (see McEnery and Hardie 2012: 8). It consists of 74,545 words of orthographic transcription of semi-structured interviews with 23 Ugandan graduates and undergraduates who speak Luganda, Runyankole-Rukiga or Luo language. The WordSmith software was used to identify the use of the 23 prepositions in the data by Luganda (29,100 words), Runyankole-Rukiga (18,610 words), and Luo (26,835 words) first language speakers and thus calculate their normalized frequencies per 1000 words.

The findings show that in the Ugandan data, the variation from Standard English in the use of preposition *in* mainly involves *in* instead of *at* as in (1) accounting for 75.53% of all variations of prepositions of denoting position.

(1) I was teaching Latin *in* Kitabi seminary.

The findings further show that the use of *in* instead of *to* is the most frequently used variation involving prepositions of destination with accounting for 44% of all cases involving variation in prepositions of destination as in (2).

(2) I was also admitted *in* Kitovu Saint Henry's.

The findings suggest that in Ugandan English the variation from Standard English in the use of preposition *in* involves mainly *at* and *to* when used as a preposition of position and destination respectively.

References

McEnery, T. & A. Hardie. 2012. *Corpus Linguistics: Methods, Theory and Practice.* Cambridge: Cambridge University Press.

Mufwene, S. 2013. Driving forces in English Contact Linguistics. In D. Schreier & M. Hundt eds. *English as a Contact Language*. Cambridge: Cambridge University Press, 204-21.

Mwangi, S. 2003. *Prepositions in Kenyan English: A Corpus-Based Study in Lexical-Grammatical Variation.* Aachen: Shaker Verlag.

Mwangi, S. 2004. Prepositions vanishing in Kenya. *English Today*, 20, 1: 27-32.

Schmied, J. 2008. East African English Kenya, Uganda, and Tanzania: morphology and syntax. In R. Mesthrie ed. *Varieties of English 4: Africa, South and Southeast Asia*. Berlin: Mouton de Gruyter, 451-71.

Introducing the corpus of *Late Modern English Medical Texts 1700—1800*

Irma Taavitsainen (University of Helsinki)
Päivi Pahta (University of Tampere) with
Turo Hiltunen, Anu Lehto, Ville Marttila, Maura Ratia, and Carla Suhr (University of Helsinki) & Jukka Tyrkkö (University of Tampere)

In this poster, we will describe the corpus of *Late Modern English Medical Texts* (LMEMT), a new resource to facilitate a systematic study of features of medical writing in the eighteenth century. Research on the language of medicine in this period has been impeded by lack of a standard corpus resource that would be sufficiently large to enable the systematic analysis of a large variety of linguistic features. Several aspects of the history of this register are still understudied and poorly understood. Yet the eighteenth century is an extremely important period in the history of medical science, representing a transfer from the earlier periods towards the more modern approaches to medicine.

LMEMT will contain c. 2 million words and reflect an inclusive view of medicine that covers the domain, including both elite and household practices. It will continue our series of two previous corpora of medical writing, *Middle English Medical Texts* 1375-1500 (MEMT, see Taavitsainen, Pahta and Mäkinen 2005) and *Early Modern English Medical Texts 1500-1700* (EMEMT, see Taavitsainen et al. 2010). The texts in LMEMT have been systematically selected in collaboration with medical historians in order to represent the wide variety of medical texts in the period. In the same way as in MEMT and EMEMT, the texts are divided into discrete categories to facilitate studies on different sub-registers of medical writing. LMEMT contains six categories: General treatises and textbooks, Texts on specific diseases, methods and midwifery, Recipe collections, Surgical and anatomical texts, Public health and Periodicals.

LMEMT will be encoded using Extensible Markup Language (XML) (Bray et al. 2008) following the principles and practices recommended in the *Guidelines for Electronic Text Encoding and Interchange* developed by the Text Encoding Initiative (TEI) Consortium (TEI Consortium 2010). Although widely adopted in digital humanities, TEI XML remains relatively rare in corpus compilation. The use of XML will facilitate easy links between the corpus texts and metadata and, later on, extending the information content of the corpus with new layers of annotation such as part-of-speech tagging and more finely developed annotation of discursive elements. At present, the markup serves to extend the expressive resources of the corpus text by representing the paratextual aspects of the printed original (Buzzetti 2009). In addition to the usual annotation of text structure (chapters, headings, paragraphs, etc.), the corpus texts will also be annotated with features relevant for research questions on book history, e.g. typographic features, lay-out and illustrations.

One of our main principles is that the corpus findings should be analysed in their sociohistorical context. For this purpose, the corpus will provide historical and sociolinguistic information about the texts and the people involved in their production. Together with the texts, LMEMT will make available a catalogue providing the complete bibliographical information of each text, a description on its subject matter as well as information about the

authors and translators. The catalogue will also cast light on the intended audience of the work and on the importance of the text. As in EMEMT, the catalogue entries are linked to images of the original texts in the *Eighteenth Century Collections Online* (ECCO) allowing the users to view and study features of typography, layout, and the relationship between images and text (subject to ECCO subscription), and also links to further authorial information in the *Oxford Dictionary of National Biography* (ODNB). Our intention is make the corpus available with its background information.

References

Bray, T., J. Paoli, C. M. Sperberg-McQueen, E. Maler & F. Yergeau. 2008. Extensible Markup Language (XML) 1.0 (Fifth Edition). W3C Recommendation 26 Nov 2008. http://www.w3.org/TR/REC-xml/.

Buzzetti, D. 2009. Digital editions and text processing. In M. Deegan & K. Sutherland eds. *Text Editing, Print and the Digital World*. Farnham: Ashgate, 45-62.

EMEMT: *Early Modern English Medical Texts* (2010) Compiled by I. Taavitsainen, P. Pahta, T. Hiltunen, V. Marttila, M. Mäkinen, M. Ratia, C. Suhr & J. Tyrkkö, with software by R. Hickey. Amsterdam: Benjamins. CD-ROM with an accompanying book: I. Taavitsainen & P. Pahta eds. *Early Modern English Medical Text: Corpus Description and Studies*.

ECCO: *Eighteenth-century Collections Online*. http://quod.lib.umich.edu/e/ecco/

LMEMT Forthcoming. *Late Modern English Medical Texts 1700-1800.* Compiled by I. Taavitsainen, T. Hiltunen, A. Lehto, V. Marttila, R. Oinonen, P. Pahta, M. Ratia, C. Suhr & J. Tyrkkö..

MEMT: *Middle English Medical Texts.* 2005. Compiled by I. Taavitsainen, P. Pahta & M. Mäkinen, with software by R. Hickey. Amsterdam: Benjamins. CD-ROM.

ODNB: Oxford Dictionary of National Biography. http://www.odnb.com.

Past tense omission in Asian Englishes: A frequency-based approach

Laura Terassa (University of Freiburg, Germany)

Terms such as "Asian Englishes" tend to imply structural similarities among geographically close varieties, although upon closer examination remarkable differences in linguistic structure can emerge (cf. Lim & Gisborne 2009). From a language contact perspective, structural diversity is explicable by distinct typological profiles in the contact-ecological setting, i.e. by substrate influence. Comparatively little attention, however, has been paid to the extent to which the frequency of use of features in the contact varieties themselves leads to variety-specific developments (cf. Bao 2010). Corpus data can help to approximate regional frequencies of use and allow for cross-varietal analyses.

This poster deals with omission of past tense marking in four Asian contact varieties of English and investigates in how far regional usage patterns can explain differences in omission rates. The varieties investigated are Hong Kong English (HKE), Singapore English (SgE), Indian English (IndE) and Philippine English (PhilE). Regional differences in the omission of past tense marking are identified by taking for each variety both the spoken parts of the International Corpus of English (ICE) and the recently released Corpus of Global Web-Based English (GloWbE; cf. Davies 2013) into account. This approach provides insights into the comparability of omission rates in data derived from the web with omission rates

observable in spoken language. The underlying assumption is that if there is a trend towards omission, omission is particularly likely to affect verbs of high absolute token frequency in both domains. This goes in line with the usage-based assumption that frequencies of use impact on the development of language structure (cf. Bybee 2007).

A logistic regression reveals that absolute token frequencies are no sufficient predictor for omission rates in the four varieties investigated. Irregular verbs are an exception in that they show considerably fewer instances of lack of past tense marking than regular verbs despite their high frequency of occurrence. This finding can be explained with the so-called "conserving effect", according to which highly frequent forms, e.g. many irregular verbs, actually resist grammatical change (cf. Bybee 2007). A comparison of these findings with estimates of the pervasiveness of past tense omission in the four varieties drawn from the electronic World Atlas of Varieties of English (eWAVE; Kortmann & Lunkenheimer 2013) shows that for each variety the eWAVE estimates by far exceed the omission rates observed in ICE and GloWbE. This might be due to the fact that the eWAVE estimates represent prototypical images of the varieties investigated. Possibly, relative rather than absolute token frequencies can explain the observed rates of past tense omission. Additionally, similar degrees of past tense omission in HKE and SgE compared to IndE and PhilE indicate that substrate influence might account for the observed differences in omission rates.

The results show that while substrate influence cannot be neglected when analyzing developments and changes in contact varieties of English (e.g. Sharma 2009), it is worth approaching cross-varietal differences from a usage-based perspective, i.e. by taking variety-specific frequency distributions into account. With regard to SgE, for instance, Low (2014: 454) stresses that there is "an urgent need for empirical validation" of theoretical findings, also in comparison with other varieties of English. This motivates a data-driven approach to variation in World Englishes.

References

Bao, Z. 2010. A usage-based approach to substratum transfer: the case of four unproductive features in Singapore English. *Language* 86(4): 792-820.

Bybee, J. L. 2007. *Frequency of Use and the Organization of Language.* Oxford: Oxford University Press.

Davies, M. 2013. *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*, http://corpus.byu.edu/glowbe/. (Accessed 25 Nov 2014.)

Kortmann, B. & K. Lunkenheimer eds. 2013. *The Electronic World Atlas of Varieties of English.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://ewave-atlas.org/. (Accessed 3 Dec 2014.)

Lim, L. & N. Gisborne. 2009. The typology of Asian Englishes: setting the agenda. *English World-Wide* 30(2): 123-32.

Low, E.-L. 2014. Research on English in Singapore. *World Englishes* 33(4): 439-57.

Sharma, D. 2009. Typological diversity in New Englishes. *English World-Wide* 30(2): 170-95.

Comparison of weekday tokens in the Brown Family of corpora

Shunji Yamazaki (Daito Bunka University)

As noted by Quirk *et al* (1985: 692), some time adverbials in English may take "the form of a noun phrase instead of a prepositional phrase". Such variation can alternatively be described as variable overt vs. zero-marking of the adverbial by a preposition, e.g. *I'll see you (on) Monday*. There are linguistic conditions under which this variation is limited: e.g. Quirk *et al* (1985: 692), Quirk and Greenbaum, (1973: 156), and Celce-Murcia and Larsen-Freeman (1999: 403) point out that prepositions of time adverbials are always absent immediately before the deictic words *last, next, this, that,* and before the quantitative words *some* and *every*. Under other conditions, such as when the adverbial denotes delimited periods of time including 'years', 'months', 'weeks', or 'days of the week', both alternatives are possible. However, several researchers have suggested that there may be context- or dialect-sensitive variation in their frequencies of use. Sonoda (2002: 19) comments that there is some conditioning by formality level: "*on* and *for* are omitted most frequently in informal styles". Algeo (1988: 14) states that with such (named) periods of time, "the omitted preposition is Common English", but that there are several areas of difference between British and American English: in some cases, British English "has no preposition, but one would be expected in American" English, and by contrast "British [English] usually requires a preposition (*on*) with days of the week, whereas American [English] can have the preposition or omit it".

The present research compares data from the Brown Family of corpora to examine the following dimensions of variation in omission of adverbial prepositions:

1. Variation by preposition (some prepositions may be more likely to be omitted than others: for example, as a function of overall preposition frequency, or (conversely) specificity of meaning).
2. Variation by regional differences (prepositions are more omitted in American English than British English).
3. Variation with sentence position of the adverbial (preposition omission may be expected with higher frequency in sentence-initial adverbials than in sentence-final adverbials).
4. Variation by genre (in particular, more formal genres should more often favour overt markers).
5. Variation by semantic relationship between the sentence and the adverbial (preposition omission should be more frequent with more general time adverbial meanings, e.g. with expressions of time duration).
6. Variation by lexical item and/or frequency (some frequent expressions may favour preposition omission).

References

Algeo, J. 1988. British and American grammatical differences. *International Journal of Lexicography*, 1: 1-31.

Celce-Murcia, M. & D. Larsen-Freeman. 1999. *The Grammar Book*. Boston: Heinle-Heinle.

Quirk, R. & S. Greenbaum. 1973. *A University Grammar of English*. Essex: Longman.

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

Sonoda, K. 2002. Omission of prepositions in time adverbials in present-day spoken AmE. *Bulletin of Nagasaki University School of Health Sciences*. 15(2): 19-25.

List of Delegates and Presenters

Carlos Acuña-Fariña	University of Santiago de Compostela
Karin Aijmer	University of Gothenburg
Dilini Algama	Justus Liebig University Giessen
Moisés Almela Sánchez	University of Murcia
Laura Altenkirch	Johannes Gutenberg Universität Mainz
Lieselotte Anderwald	University of Kiel
Lucija Antolkovic	Uppsala University
Sabine Arndt-Lappe	University of Trier
Yinchun Bai	Uni Freiburg / Uni Antwerpen
Sabine Bartsch	Technische Universität Darmstadt
Cynthia Berger	Georgia State University
Erika Berglind Söderqvist	Uppsala University
Eva Berlage	Universität Hamburg
Tobias Bernaisch	Justus Liebig University Giessen
Yves Bestgen	University of Louvain
Carolin Biewer	University of Bonn
Susan Blackwell	Language Consultancy Desk
Cristina Blanco-García	Universidad de Santiago de Compostela
Anne-Katrin Blass	University of Trier
Lieselotte Brems	University of Liège / KU Leuven
Marisa Brook	University of Toronto
Elisabeth Bruckmaier	LMU Munich
Marie-Louise Brunner	Saarland University
Sara Budts	KU Leuven
Kate Burridge	Monash University
Beatrix Busse	University of Heidelberg
Elena Callegaro	University of Zurich
Marcus Callies	University of Bremen
Anna Cermáková	Charles University
Lucie Chlumská	Charles University
Claudia Claridge	University of Duisburg-Essen
Caroline Collet	Saarland University
Eduardo Coto Villalibre	University of Cantabria
Kristin Davidse	KU Leuven
Sylvie De Cock	Université catholique de Louvain
Rachele De Felice	University College London
Johan de Joode	University of Nottingham
Sandra Deshors	New Mexico State University
Stefan Diemer	Saarland University
Hildegunn Dirdal	University of Oslo
Florian Dolberg	Johannes Gutenberg-Universität
Steffi Dose-Heidelmayer	
Maïté Dupont	Université catholique de Louvain
Signe Oksefjell Ebeling	University of Oslo
Alison Edwards	University of Cambridge
Thomas Egan	Hedmark University College
Matthias Eitelmann	Johannes Gutenberg-University Mainz
Stefan Evert	FAU Erlangen-Nürnberg
Teresa Fanego	University of Santiago de Compostela
Yolanda Fernández-Pena	University of Vigo
Roswitha Fischer	University of Regensburg
Susanne Flach	Freie Universität Berlin
Robert Fuchs	Westfälische Wilhelms-Universität Münster
Costas Gabrielatos	Edge Hill University
Gregory Garretson	Uppsala University
Matt Gee	Birmingham City University
Lobke Ghesquière	FWO / KU Leuven / Vrije Universiteit Brussel
Gaëtanelle Gilquin	FNRS / Université catholique de Louvain
Sandra Götz	Justus Liebig University Giessen
Sylviane Granger	University of Louvain
Melanie Green	University of Sussex
Stefan Th. Gries	University of California, Santa Barbara
Barbara Ann Güldenring	Philipps-Universität Marburg
Natalia Gvishiani	Moscow Lomonosov State University
Naomi Hallan	University of Trier

Beke Hansen	University of Kiel
Andrew Hardie	Lancaster University
Hilde Hasselgård	University of Oslo
Abi Hawtin	Lancaster University
Benedikt Heller	KU Leuven
Thomas Herbst	FAU Erlangen-Nürnberg
Thomas Hoffmann	KU Eichstätt-Ingolstadt
Knut Hofland	Uni Research Computing, Bergen
Mikko Höglund	Stockholm University
Stephanie Horch	University of Freiburg
Magnus Huber	University of Giessen
Jennifer Hughes	Lancaster University
Marlén Izquierdo	University of the Basque Country EHU/UPV
Bridget Jankowski	University of Toronto
Berit Johannsen	Freie Universität Berlin
Amelia Joulain-Jay	Lancaster University
Mark Kaunisto	University of Tampere
Andrew Kehoe	Birmingham City University
John Kirk	Technische Universität Dresden
Juhani Klemola	University of Tampere
Gabriele Knappe	University of Bamberg
Róisín Knight	Lancaster University
Christopher Koch	University of Technology Dresden
Thomas Kohnen	Universität zu Köln
Daniela Kolbe-Hanna	University of Trier
Rolf Kreyer	Uni Marburg
Haidee Kruger	Macquarie University / North-West University
Kostiantyn Kucher	Linnaeus University
Gero Kunter	Heinrich-Heine-Universität Düsseldorf
Kerstin Kunz	University of Heidelberg
Kristopher Kyle	Georgia State University
Merja Kytö	Uppsala University
Mikko Laitinen	Linnaeus University
Alexander Lakaw	Linnaeus University
Samantha Laporte	Université catholique de Louvain
Ekaterina Lapshinova-Koltunski	Universität des Saarlandes
Tove Larsson	Uppsala University
Jacqueline Laws	University of Reading
Marie-Aude Lefer	Marie Haps School of Translation and Interpreting
Hans Martin Lehmann	University of Zurich
Magnus Levin	Linnaeus University
Signe-Anita Lindgrén	Abo Akademi University
María José López-Couso	University of Santiago de Compostela
David Lorenz	Albert-Ludwigs-Universität Freiburg
Lucía Loureiro-Porto	University of the Balearic Islands
Robbie Love	Lancaster University
Kerstin Lunkenheimer	University of Trier
Markéta Malá	Charles University in Prague
Sebastian Malinowski	Saarland University
José Manuel Martínez Martínez	Universität des Saarlandes
Beatriz Mato-Míguez	Universidade de Santiago de Compostela
Willem Meijs	Language Consultancy Desk (LCD)
Belén Méndez-Naya	University of Santiago de Compostela
Katrin Menzel	Universität des Saarlandes
Maja Milıčević	University of Belgrade
Ilka Mindt	Universität Paderborn
Ruth Moehlig-Falke	University of Heidelberg
Susanne Mohr	University of Bonn
Alessandra Molino	University of Turin
Britta Mondorf	University of Mainz
Lilo Moessner	RWTH University Aachen
Lilo Mössner	RWTH University Aachen
Danny Mukherjee	Justus Liebig University Giessen
Akira Murakami	University of Birmingham
Taichi Nakamura	Senshu University
Stella Neumann	RWTH Aachen University
Ngum Meyuhnsi Njende	Katholieke Universiteit Leuven
Paloma Núñez Pertejo	University of Santiago de Compostela
Arja Nurmi	University of Tampere
Gloria Otchere	University of Oslo
Gabriel Ozon	University of Sheffield
Päivi Pahta	University of Tampere
Ignacio Palacios	University of Santiago de Compostela

Magali Paquot	FNRS - Université catholique de Louvain
Hanna Parviainen	University of Tampere
Pam Peters	Macquarie University
Peter Petré	Université de Lille 3 (UMR8163) / KU Leuven
Michael Pidd	University of Sheffield
Marie-Luise Pitzl	University of Vienna
Isabel Pizarro	Universidad de Valladolid
Nele Põldvere	Lund University
Rosa Rabadán	University of León
Paula Rautionaho	University of Tampere
Paul Rayson	Lancaster University
Antoinette Renouf	Birmingham City University
Paula Rodriguez	University of Cantabria
Katja Roller	University of Freiburg
Ute Römer	Georgia State University
Patricia Ronan	University of Lausanne
Sylvi Rørvik	Hedmark University College
Anna Rosen	University of Freiburg
Christoph Rühlemann	University of Paderborn
Tanja Rütten	University of Cologne
Tanja Säily	University of Helsinki
Christina Sanchez-Stockhammer	Universität Erlangen-Nürnberg
Andrea Sand	University of Trier
Simon Sauer	Humbolt-Universität zu Berlin
Teri Schamp-Bjerede	Lund University
Julia Schlüter	University of Bamberg
Christa Schmidt	RWTH Aachen University
Selina Schmidt	Saarland University
Gerold Schneider	University of Zurich
Ulrike Schneider	University of Mainz
Verena Schröter	Albert-Ludwigs-Universität Freiburg
Alison Sealey	Lancaster University
Jess Shapero	Freelance English Language Consultant
Lucia Siebers	University of Regensburg
Jolanta Sinkūnienė	Vilnius University
Urszula Skrzypik	University of Chester
Adam Smith	Macquarie University
Gillian Smith	Lancaster University
Nick Smith	University of Leicester
Audronė Solienė	Vilnius University
Myounghyoun Song	Seoul National University
Susana Sotillo	Montclair State University
Jude Ssempuuma	Ruhr University Bochum
Ulrike Stange	Johannes Gutenberg-Universität Mainz
Helene Steigertahl	Universität Bayreuth
Erich Steiner	Universität des Saarlandes
Jenny Ström Herold	Linnaeus University
Michael Stubbs	University of Trier
Cristina Suárez-Gómez	University of the Balearic Islands
Irma Taavitsainen	University of Helsinki
Sali Tagliamonte	University of Toronto
Iván Tamaredo	University of Santiago de Compostela
Masayuki Tamaruya	Rikkyo University
Laura Terassa	University of Freiburg
Paul Thompson	University of Birmingham
Shin'ichiro Torikai	Rikkyo University
Gunnel Tottie	The University of Zurich
Graeme Trousdale	University of Edinburgh
Jukka Tuominen	University of Tampere
Jukka Tyrkkö	University of Tampere
C U C Ugorji	University of Benin
Peter Uhrig	FAU Erlangen-Nürnberg
Aurelija Usoniene	Vilnius University
Kees Vaes	John Benjamins
Eric Van Broekhuizen	Brill academic publishers
Bertus Van Rooy	North-West University
Turo Vartiainen	University of Helsinki
Ake Viberg	Uppsala University
Sebastian Wagner	University of Duisburg-Essen
Stephen Wattam	Lancaster University
Edmund Weiner	Oxford University Press / OED
Martin Weisser	Guangdong University of Foreign Studies
Bianca Widlitzki	Justus Liebig University Giessen
Viola Wiegand	The University of Nottingham
Leonie Wiemeyer	University of Bremen
Christoph Wolk	Justus-Liebig-Universität Gießen
Shunji Yamazaki	Daito Bunka University
Nuria Yáñez-Bouza	Universidade de Vigo
Ekaterina Zaytseva	University of Bremen