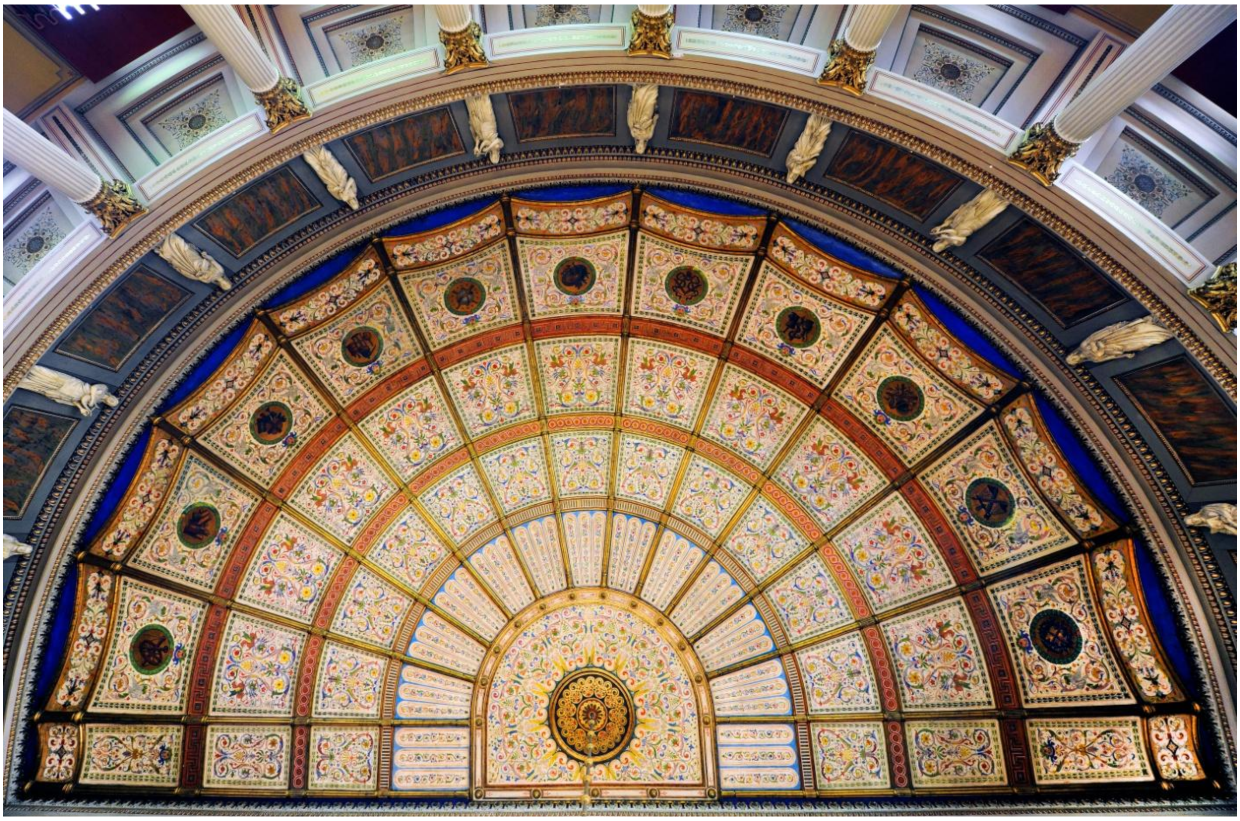
ICAME 32

Oslo 2011

Trends and Traditions in English Corpus Linguistics
In Honour of Stig Johansson

ABSTRACTS



http://www.uio.no/icame2011

ICAME 32 – Trends and Traditions in English Corpus Linguistics
In Honour of Stig Johansson (1939-2010)

It is in deep gratitude and with great respect that we dedicate this year's ICAME conference to the memory of Stig Johansson, who passed away on 22 April 2010. Stig was a founder member of ICAME and served as its first coordinating secretary for many years. Among his achievements are the completion of the LOB corpus, participation in the Text Encoding Initiative, development and compilation of the English-Norwegian Parallel Corpus and co-authorship of the *Longman Grammar of Spoken and Written English.* While Stig represented ICAME tradition, he also saw change as a vital part of corpus linguistics, looking ahead with entrepreneurial spirit. It was his idea to get the ICAME conference to Oslo, and he took part in its early preparations. He also donated his remaining research funds to the conference.

Stig's academic and personal merits are too many to be listed in this brief note, but one can get a glimpse at the website set up for his 70th birthday at http://gandalf.uib.no/Stig. In a thank-you message to all those who had contributed to the website, Stig summed up his principle of research, with which we will end this brief note, as it was typical of Stig to share his experience and offer advice to his students and colleagues:

> I take this opportunity to add my principle for research which I have (tongue-in-cheek) christened the IKEA principle, since I come from the same little corner of Småland as the IKEA man, Ingvar Kamprad. The elements are in this order:
>
> | | |
> |---|---|
> | **I** | idéer, ideas |
> | **K** | kunnskap, knowledge |
> | **E** | entusiasme, enthusiasm |
> | **A** | arbete, work |
>
> Notice that I did not mention money. Money helps of course, but without the other elements it may accomplish very little.
>
> (Stig Johansson, March 2009)

From the first ICAME seminar, Bergen, March 1979. From the left: Jostein Hauge, Director of the Norwegian Computing Centre for the Humanities, W. Nelson Francis, Geoffrey Leech, Stig Johansson, Arne Zettersten, Henry Kučera, Randolph Quirk and Jan Svartvik

Contents

Acknowledgements

We warmly thank the following sponsors and organisations for their kind support.

UiO : University of Oslo

The Research Council of Norway

NHH 2011 75 år

UNIVERSITAS BERGENSIS
UNIVERSITETET I BERGEN

JOHN BENJAMINS PUBLISHING COMPANY
www.benjamins.com

uni Research

Edinburgh University Press
THE UNIVERSITY OF EDINBURGH

CAMBRIDGE UNIVERSITY PRESS

ROUTLEDGE
Routledge Taylor & Francis Group

continuum

Rodopi

PETER LANG
INTERNATIONAL ACADEMIC PUBLISHERS

Organising committee
Hilde Hasselgård (chair, Oslo)
Kristin Bech (Oslo)
Jarle Ebeling (Oslo)
Signe Oksefjell Ebeling (Oslo)
Gisle Andersen (NHH, Bergen)
Knut Hofland (Uni Research AS, Bergen)

Student helpers
Ada Benedicte Aydin
Siri Heslien
Torunn Johansen
Hege Larsson Aas
Elisabeth Maria Neuhaus

Front page photography of the ceiling of the Old Ceremonial Theatre in the University's *Domus Academica* building: Arthur Sand (UiO)

Plenary speakers

‘O brave new world, that has such corpora in it!’ New trends and traditions on the Internet

David Crystal
University of Bangor

Electronically mediated communication is fundamentally altering our notion of text and text type, and presents corpus linguistics with a fresh set of challenges. There are some continuities with traditionally spoken and written text, but also important discontinuities. A pragmatic perspective brings to light new kinds of text, such as those which include features to defeat spam filters or to ensure a high search-engine ranking, those which are the product of specific technological constraints (as in Twitter), and those which raise ethical or legal issues, especially in areas of sensitivity such as online security and paedophile activity. Specific problems for current corpus linguistics include how to handle written texts whose boundaries are continually changing or where authorship is unclear, as in contexts where there is moderation or interactivity. The main issue for the future will be how to deal with the increased presence of spoken texts, as a result of growth in Voice over the Internet and mobile communication. Whatever the trends and traditions were in corpus linguistics during its first half century, they will be very different in its next.

Missing connectives – and what they may tell us

Cathrine Fabricius-Hansen
University of Oslo

Prepositions, connectives, and other ‘functional’ words are very suitable objects for contrastive studies based on parallel corpora like the ENPC and its extension, the Oslo Multilingual Corpus: They are easily searchable, even without PoS tagging, and they tend to be fairly frequent in running text so that even small-scale corpora may yield interesting and robust results, as demonstrated in a number of studies by Stig Johansson and Bengt Altenberg, among others.

One of the intriguing observations concerning connectives (in a broad sense) is that sometimes they are left out in translations or, conversely, added ‘out of the blue’ even when the target, or source, language has an adequate explicit expression at its disposal. Such ‘missing connectives’ – or zero correspondences – are the topic of my talk.

I shall first present results from earlier OMC-based case studies indicating that German uses certain adverbial connectives more extensively than Norwegian and, in particular, English. I then go on to discuss more in detail possible interpretation effects of adding or leaving out these connectives, and possible explanations for the zero ratio differences observed in the data. I shall argue that corpus-based hypotheses in this area should be checked experimentally.

Continuity and discontinuity between early and late Middle English – comparing the maps in LAEME and e-LALME

Margaret Laing
University of Edinburgh

A Linguistic Atlas of Early Middle English (*LAEME*) was published as a free-access website in 2008. Although much of the work is thus available, it was not 'complete' at the time of first publication, additions and corrections being periodically made to it since, including a number of maps showing early Middle English dialectal features. More are forthcoming. Moreover, between September 2007 and August 2010, Keith Williamson and I at the Institute for Historical Dialectology, University of Edinburgh, in collaboration with Michael Benskin, of the University of Oslo were engaged on an AHRC-funded project to produce **a revised on-line edition of** *A Linguistic Atlas of Late Mediaeval English* (e-*LALME*). The provision of on-line maps formed part of the enhancement project. By the time of ICAME 32, it should therefore be possible to view online sets of maps from both *LAEME* and e-*LALME* in order to compare regional distributions of forms between the two periods covered by these atlases.

Far less source material survives from 1150–1325 than from 1350–1450 and the regional coverage provided by *LAEME* is less dense and very patchy compared to that of *LALME*. Nevertheless, for the first time it will be possible to compare regional distributions of different features appearing in forms for the same linguistic item from two contiguous periods of the history of English. In this presentation, I hope to be able to illustrate three types of distributions:

(a) those where the regional pattern of selected features in the attested forms for a particular item is largely continued from early Middle English into late Middle English
(b) those showing how some features may be recessive and some emergent between early Middle English and late Middle English
(c) those that illustrate a comparatively mixed picture for early Middle English, while the late Middle English picture shows a clearer regional 'consensus'.

Writing the corpus-based history of spoken English

Christian Mair
University of Freiburg

The past two decades have seen considerable advances in the real-time corpus-based investigation of linguistic change in English, both in older stages of the language and in progress now. Closer inspection of the relevant findings, however, reveals that most claims about changes in the language as a whole have been based largely on written data. Against this backdrop, the present paper seeks to define the potential and limitations of the corpus-based study of *change in the spoken language*, where even for English the major problem is still the relative dearth of suitable corpus data.

In the absence of suitable recordings, the study of real speech in real time will never be pushed back further than the early 20thcentury, but as I will make clear with a number of examples (DCPSE, ONZE, DECTE, WW I *Lautkommission* recordings), a number of interesting resources deserve more corpus-linguistic attention than they have received so far. Considerable progress is also likely in the study of the history of the spoken language by "proxy", i.e. through speech-like genres, of which vast amounts have recently been made available for corpus-linguistic study (Old Bailey, EEBO, Google Ngrams). Informal and speech-like styles of interaction that have developed in certain domains of the World-Wide Web represent an additional promising window on very recent developments.

I will illustrate the argument with three empirical case studies: (i) the history of cleft sentences in Standard English from c. 1600 to the present, (ii) the emergence and spread of *do*-support with *got (to)/gotta* in non-standard North American English, and (iii) the mediated globalisation of vernacular features in diasporic web-forums.

Sequence and order: the neo-firthian tradition in corpus semantics

Michael Stubbs
University of Trier

Firth, Halliday, Palmer and Sinclair distinguish between "sequence" (an observable feature of texts) and "order" (a feature of linguists' models).

1. AN INITIAL EXAMPLE. In an utterance such as the following, the words "went and" can hardly be interpreted literally.

... it was all going well ... but then he *went and* spoiled it all ...

They are a conventional way of expressing surprise and/or disapproval. I will use such examples to illustrate a model of phrasal units in English.

2. INDUCTION? In building their models, corpus linguists often attempt to avoid assumptions imported from pre-corpus studies, by using methods which could be called "inductive", in so far as they proceed from observations about texts to generalizations. However, induction has been questioned for over 400 years (by Bacon, Hume, Popper and others), and few people now believe in the possibility of rigorous, theory-free induction (Black 1967).

3. SINCLAIR'S MODEL OF EXTENDED LEXICAL UNITS. Sinclair's phrasal model (1996 and later articles) is certainly not a purely inductive generalization from observed data. This model, and especially its component "semantic prosody", immediately attracted widespread interest, but also caused confusion: perhaps because critics attempted to discuss semantic prosody independently of the model of which it is only one component. I will try to make explicit the logical relations within the model, between form (lexical and syntactic) and function (semantic and pragmatic).

This model exemplifies one of the most important kinds of intellectual progress. It succeeds in relating things (lexis, syntax, semantics and pragmatics) which were previously poorly related, in revealing connections between different areas of an academic field, and in introducing order where there was previously disorder.

4. PROBLEMS FOR RESEARCH. Several aspects of the phrasal model certainly remain unclear: e.g. whether it can be rigorously tested, or indeed how many such phrasal units there are (in English). As examples of other aspects of the model which require to be developed, I will dissuss:

(a) the functions of phrasal units in textual organization
(b) the relations between paraphrase, intertext and meaning.

References

Max Black (1967) Induction. In P Edwards ed *The Encyclopedia of Philosophy*. NY: Macmillan. 169-81.

John McH Sinclair (1996) The search for units of meaning. *Textus*, 9, 1: 75-106. [Reprinted in J Sinclair (2004) *Trust the Text*. Routledge. 24-48.]

Pre-conference workshops

Workshop 1: Corpus-based contrastive analysis

Karin Aijmer, University of Gothenburg
Bengt Altenberg, Lund University

In 1993, at ICAME 14 in Zürich, Stig Johansson presented a corpus-based project that was to begin a new era in contrastive linguistics and translation studies. Using a computer corpus of comparable English and Norwegian texts and their translations into the other language, he and his team created a fruitful empirical basis for comparing the systems and use of the two languages from lexis to discourse. Since then, the idea of using bilingual or multilingual translation corpora (together with comparable corpora of original texts for control purposes) has spread and a number of researchers are now using this approach to compare different sets of languages and to develop methodologies for various practical applications offered by the corpora, e.g. in language teaching, lexicography, machine-aided translation, and automatic lexicon extraction.

The purpose of the workshop is to bring together researchers involved in the use of bilingual or multilingual corpora for various purposes, theoretical or practical, to exchange views and experiences and, not least, to get to know each other.

Using recurrent word-combinations to explore cross-linguistic differences

Jarle Ebeling, Signe O. Ebeling, Hilde Hasselgård
University of Oslo

In this study we will investigate phraseological differences between English and Norwegian using a balanced, bidirectional parallel corpus, viz. The English-Norwegian Parallel Corpus. If it is the case, as many linguists have suggested, that meaning in text is rarely expressed by single-word items, should not cross-linguistic studies (of text) take multi-word units as their point of departure?

To bootstrap our investigation lists of recurrent three-word combinations with a frequency of at least 10 in English and Norwegian original and translated fiction texts were produced using AntConc (http://www.antlab.sci.waseda.ac.jp/antconc_index.html). The hypothesis was that such lists could point to cross-linguistic differences that might elude corpus investigations that take specific lexicogrammatical constructions as their starting point. Three-word combinations were chosen as they have been shown to be frequent enough to yield interesting data (Altenberg 1998) and were thought to reveal meaningful patterns.

Examples of differences that come to light include the following word-combinations in English and Norwegian:

Three-word combination	No. of occurrences in English original texts	No. of occurrences in English translated texts	No. of occurrences in Norwegian original texts	No. of occurrences in Norwegian translated texts
a long time	23	103		
the same time	18	76		
for a moment	37	79		
all the way	17	49		
all the same	13	37		

i det hele	(Gloss: *on the whole* < 10)	(Gloss: *on the whole* < 10)	34	51
i det minste	(Gloss: *in the least* < 10)	(Gloss: *in the least* < 10)	20	27

The next step in the analysis is to find out what triggers or lead to these quantitative differences between original and translated text or between English and Norwegian. Do the divergences point to systematic differences between the two languages, or can they be explained by concepts such as translationese (Gellerstam 1986, Johansson 2007: 32 f) and translation strategies (Baker 1992: 26 ff)?

In addition to seeking the answer to questions such as the ones above, the paper will discuss methodological issues related to using recurrent word-combinations as a point of departure for contrastive studies and more generally, what we mean by phraseological differences.

In this study we wish to honour Stig Johansson in exploring new avenues in cross-linguistic research by using a bidirectional parallel corpus in phraseologically oriented contrastive studies.

References

Altenberg, Bengt. 1998. On the Phraseology of Spoken English: The Evidence of Recurrent Word-combinations. In A.P. Cowie (ed.). Phraseology. Theory, Analysis, and Applications. Oxford University Press. 101-122.

Baker, Mona. 1992. In Other Words. A Coursebook on Translation. London and New York: Routledge.

Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist (eds.), Translation Studies in Scandinavia. Lund: CWK Gleerup. 88-95.

Johansson, Stig. 2007. Seeing through Multilingual Corpora. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sinclair, John McH. 1996. The search for units of meaning. Textus, 9/1: 75-106. [Reprinted in J. Sinclair (2004) Trust the Text. Routledge. 24-48.]

Tertia Comparationis in Multilingual Corpora

Thomas Egan
Hedmark University College

Johansson (1997:39) touches on the status of *tertium comparationis* in contrastive studies. He writes: “Much discussion in contrastive analysis has revolved around the question of the *tertium comparationis*, i.e. the background of sameness against which differences can be viewed and described”. However, the status of various sorts of *tertia comparationis* would seem to have been more of a topic of discussion among pragmatists and sociolinguists than among corpus linguists (see references in Jaszczolt 2003). Nevertheless, any contrastive corpus linguist who takes translation equivalence as evidence of semantic equivalence is working on the overt or tacit assumption that there exists a viable *tertium comparationis* in the form of a meaning component common to both the source expression and its translation.

One problem with using translation equivalence as a *tertium comparationis*, according to Krzeszowski (1990: 18), lies in the distinction between semantic and pragmatic content. Another, related problem is the delimitation of what is taken to be the semantic or pragmatic content common to the two expressions. Krzeszowski (1990:25) employs the term *2-text* to refer to texts in either parallel or translated corpora. The availability of multilingual corpora, such as the Oslo Multilingual Corpus, allows us to operate with the concept of the *3-text*, with expressions in a source language serving as potential *tertia comparationis* for their translations into two other languages. This means that we can bypass (or beg?) the second problem raised by Krzeszowski, the identification of semantic content common to a source item and its translation. What the two sets of translated items have in common is simply the fact that they are both translations of the same source items. As for the first problem, establishing the borderline between semantics and pragmatics, this rests on a distinction between the two that is increasingly seen as arbitrary and unwarranted (see, for example, Langacker 2008: 40).

In this paper I operationalise the notion of *tertium comparationis* in *3-text*s in a study of how the two notions of *betweenness* and *throughness* are encoded in English and French, comparing translation equivalents of the Norwegian prepositions *mellom*, which encodes the *betweenness* relationship, and *gjennom*, which encodes the *throughness* relationship. All tokens of the two Norwegian prepositions in the Oslo Multilingual Corpus were classified according to the semantic type of predication encoded by the preposition. The Norwegian originals were then set aside and comparisons drawn between the English and French renderings of the various meanings. Statistical calculations were employed to establish whether the forms of translation of the various semantic classes differ significantly from those of the other classes, both within English and French and across the two languages.

References

Jaszczolt, K. M. (2003) 'On translating what is said: *tertium comparationis* in contrastive semantics and pragmatics'. In: K. M. Jaszczolt and K. Turner (eds). *Meaning Through Language Contrast*. Amsterdam: John Benjamins. Vol. 2. 441-462.

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

Krzeszowski, Tomasz P. (1990) Contrasting languages: the scope of contrastive linguistics. Berlin: Mouton de Gruyter

Langacker, R. (2008) *Cognitive Grammar: A basic introduction*. Oxford: Oxford University Press.

Enriching the phraseological coverage of bilingual dictionaries: the respective contribution of monolingual and bilingual corpus data

Sylviane Granger, Marie-Aude Lefer
Catholic University of Louvain

As pointed out by Stig Johansson (2007: 308), one of the most obvious applications of multilingual corpus research is in bilingual and multilingual lexicography. One of the aspects of bilingual dictionaries that has benefited most from corpus analysis is their phraseological coverage (Lubensky & McShane 2007, Ferraresi *et al.* 2010). However, the treatment of word combinations in bilingual dictionaries still lags far behind that displayed in monolingual dictionaries, a fact that was noted by Rundell (1999): “The extraordinary range of lexical and grammatical information they include is rarely even approached by the best bilingual dictionaries available”. This situation is largely due to the more limited use made of corpora in bilingual lexicography, itself due to the lack of large balanced bilingual corpora, in particular translation corpora.

In our presentation we focus on some highly frequent adverbs in English and French and compare the coverage of their multi-word uses in bilingual dictionaries with their corpus-attested usage patterns. For the dictionary analysis, we make use of three French<>English electronic dictionaries (*Le Robert & Collins, Harrap's Unabridged Pro* and *Hachette Oxford*). For the corpus investigation, we resort to a range of corpus data: monolingual corpus data (mainly the *British National Corpus*), unidirectional translation corpus data (the *Label France* corpus, a 1-million word French to English translation corpus) and bidirectional corpus data (the 1.3-million word *PLECI* corpus, which contains both fiction and journalese) as well as the *Europarl* corpus in the *Sketch Engine* (Koehn 2005, Kilgarriff *et al.* 2004) and web-based tools such as *WeBiText*[1]. We start by analyzing the highly polysemous French adverb *encore* (Mosegaard Hansen 2002) and its translations and then reverse the perspective and investigate the English adverbs that are most frequently found as translations of *encore* in the corpus, i.e. *still, again, yet* and *even*.

The study shows that the phraseological coverage of these words is quite limited and that the phraseological units, when included, tend to be poorly translated. Our study therefore confirms that "dictionaries fall short in the light of the evidence from bilingual corpora" (Johansson 2007: 308). One of the encouraging results of our study is that, in the absence of large balanced bilingual corpora, a systematic use of monolingual corpora can already go a long way towards 'phrasing up' the bilingual dictionary.

[1] http://www.webitext.com/bin/webinuk.cgi

References

Ferraresi A., S. Bernardini, G. Picci & M. Baroni (2010). Web corpora for bilingual lexicography: a pilot study of English/French collocation extraction and translation. In R. Xiao (ed.) *Using corpora in contrastive and translation studies*. Newcastle upon Tyne: Cambridge Scholars Publishing, 337-359.

Johansson, S. (2007). Seeing through multilingual corpora. On the use of corpora in constrastive studies. Amsterdam & Philadelphia: Benjamins.

Kilgarriff, A., P. Rychly, P. Smrz & D. Tugwell (2004). The Sketch Engine. In *Proceedings of Euralex 2004*, Lorient (France), 105-116. http://www.sketchengine.co.uk/

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit 2005.*

Lubensky S. & M. McShane (2007). Bilingual phraseological dictionaries. In H. Burger, D. Dobrovol'skij, P. Kühn & N.R. Norrick (eds.) *Phraseology. An International Handbook of Contemporary Research*. Berlin & New York: de Gruyter, 919-928.

Mosegaard Hansen, M.-B. (2002). La polysémie de l'adverbe encore. *Travaux de linguistique* 44: 143-166.

Rundell, M. (1999). Dictionary use in production. *International Journal of Lexicography* 12.1: 35-53.

The internal gradience of the adposition category: some evidence from comparable corpora of English, Nepali and Russian

Andrew Hardie
Lancaster University

Previous research on the Nepali language (Hardie 2008) indicates that its postposition category is *internally gradient* in terms of the predominant combinatory behaviour of the words within it, as indicated by patterns across statistical collocation data. Subsequent findings (Hardie and Mudraya 2009) suggest that, for a subset of adpositions (locative and ablative), this gradience is cross-linguistically consistent across three related languages, namely English, Nepali and Russian. The present paper extends this analysis to the full range of highly frequent adpositions in each language.

The method of analysing combinatory behaviour applied here, dubbed a *"quantitative-distributional"* approach, posits that collocational patterns observed across words within a grammatical category can be used as a means to identify defining characteristics of that category. This methodology is based on statistical measures of collocation in text corpora. First, tables of collocates for a number of search-terms are assembled – here, the twenty most significant collocates within two words of the node, based on Z-score. Then, grammatical and/or semantic patterns are identified across the collocation tables. These patterns are used to characterise commonalities across the search nodes. This corpus-based methodology is compatible with a diverse range of theoretical positions, including Hoey's (2005) Lexical Priming and Construction Grammar models (e.g. Croft 2001).

Applying this approach to three one-million-word comparable written corpora across the three languages allows two major patterns to be identified which characterise the adposition category. These are (1) co-occurrence with semantically congruent nouns; (2) co-occurrence with lexical items for which the adposition functions as a subcategoriser. Some minor co-occurrence patterns correlate with these major phenomena (e.g. in English co-occurrence with the definite article correlates with the semantic congruence pattern). These patterns are cross-linguistically consistent, although there are few or no cross-language similarities at the level of individual collocates – largely what we should expect given that much phraseology is conventional in nature. Moreover all three languages show a graded scale between adpositions where semantic congruence predominates and adpositions where subcategorisation predominates. But the arrangement of particular case functions across this "landscape" is language-particular, although not wholly arbitrary. In particular, the positions in their respective landscapes of the English dative and Nepali genitive adpositions result in part from their interaction with these languages' verbal inflection systems. Moreover, the interaction of adpositional marking with inflectional case marking (and, to a lesser extent, word order) is distinct in each language, a factor which differentiates English adpositions from Nepali and Russian adpositions.

References

Croft, W. (2001) Radical Construction Grammar: Syntactic theory in typological perspective. Oxford: Oxford University Press.

Hardie, A (2008) A collocation-based approach to Nepali postpositions. In: *Corpus Linguistics and Linguistic Theory* 4(1): 19-62.

Hardie, A and Mudraya, O (2009) Collocational patterning in cross-linguistic perspective: adpositions in English, Nepali, and Russian. *Arena Romanistica* 4: 138-149.

Hoey, M. (2005) *Lexical Priming*. London: Routledge.

German-English contrasts in cohesion

Kerstin Kunz, Erich Steiner
Saarland University

Over the last decades, substantial insights have been gained into the linguistic level of cohesion in German (see e.g. Linke, A., Nussbaumer, M. and P.R. Portmann. 2004) and, to an even greater extent, in English (e.g. Halliday & Hasan 1976, Schubert 2008, Esser 2009). However, these are mainly monolingual in coverage and example-based methodologically. In recent years, cohesive devices and their functions have been investigated empirically by conducting monolingual analyses or by contrasting English and German texts (Becher et al. 2009, Bosch et al. 2007, Gundel et al. 2004). These corpus-linguistic efforts are usually limited to the analysis of a restricted set of individual phenomena, usually in one register, and hence not widely generalizable.

Our research project, funded by the German research Foundation (DFG), aims at a contrastive model for cohesion English-German, drawing on a comparison of a broad range of systemic resources as well as their instantiations in English and German texts (for preliminary attempts cf. Hansen-Schirra et al. 2007, Kunz 2010). The combination of hermeneutic and example-based research with the corpus-linguistic analysis of English and German original texts and translations of written and spoken registers provides a broadened perspective in terms of the range of cohesive phenomena in the two languages. The corpus-linguistic approach, in particular, identifies the functions of these phenomena in different contexts, tracing co-occurrences as well as measuring frequencies of register-specific use. Furthermore, it permits comprehensive investigations of cohesive chains in terms of frequency, size, distance, function, etc.

The first part of the talk will introduce our concept of cohesion as a set of categories suitable for the comparison of cohesion English-German. This will be followed by an initial overview of systemic differences in terms of forms and functions/meanings expressed. A selection of findings from our corpus linguistic analysis will be presented in the following areas:

- Demonstrative reference: demonstrative pronouns
- Substitution: nominal
- Cohesive conjunctions: causal conjunctions

One of the long-term questions we intend to address with our research is whether contrastive properties of cohesion in the two languages point into the same direction as some assumed generalizations in contrastive grammar (e.g. more/less explicit encoding of functions/meanings via a greater/smaller range of forms or a larger/smaller number of elements in cohesive chains; more/less variation in the use of different cohesive devices), or whether cohesion serves as a dialectic counterpart, distributing constraints in different directions from those formulated for contrastive grammars of the two languages (e.g. Hawkins 1986; König und Gast 2009).

References

Bosch P., Katz G. and C. Umbach. 2007. The non-subject bias of German demonstrative pronouns. In: Schwarz-Friesel M., Consten M. and M. Knees, eds. 2007. *Anaphors in Text. Cognitive, formal and applied approaches to anaphoric reference*. Universität Jena: Studies in Language Companion Series. 145-164.

Becher V., Höder S. and S. Kranich. 2009. "A tentative typology of translation-induced language change". Paper given at Workshop "Multilingual Discourse Production", 6–7 November 2009, University of Hamburg, Research Centre on Multilingualism, Hamburg

Esser J. 2009. *Introduction to English Text-linguistics*. Frankfurt am Main: Peter Lang

Gundel J. K., Hedberg N. and R. Zacharski. 2004. "Demonstrative pronouns in natural discourse". In: *Proceedings of the Fifth Discourse Anaphora and Anaphora Resolution Colloquium*. São Miguel, Portugal. pp. 81-86.

Halliday M.A.K. and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Hansen-Schirra S., Neumann S. and E. Steiner 2007. "Cohesion and Explicitation in an English-German Translation Corpus". In: *Languages in Contrast* 7:2. 2007. pp. 241-265.

Hawkins J. A. 1986. A Comparative Typology of English and German. Unifying the Contrasts. London: Croom Helm.

König E. and V. Gast. 2009. Understanding English-German Contrasts. Grundlagen der Anglistik und Amerikanistik. Berlin: Schmidt.

Kunz K. 2010. English and German Nominal Coreference. A Study of Political Essays. Frankfurt am Main: Peter Lang.

Linke A., Nussbaumer M. and P.R. Portmann. 2004. *Studienbuch Linguistik*. 5th edition. Tübingen: Niemeyer.

Schubert C. 2008. *Englische Textlinguistik. Eine Einführung*. Berlin: Erich Schmidt.

Thematic variation in English and Spanish newspaper genres: A contrastive corpus-based study

Julia Lavid, Jorge Arús, Lara Moratón
Universidad Complutense de Madrid

This paper describes the thematic variation observed in two newspaper genres – news reports and commentaries – in English and Spanish, and studies the influence of genre-specific and language-specific constraints on the observed variation. The motivation for the study lies in the renewed interest in the contrastive analysis of certain newspaper genres such as commentaries (Wang 2008), opinion columns (Dafouz 2008) or editorials (Alonso Belmonte coord. 2007, Lavid et al. in press, Tirkkonen-Condit 1996, *inter alia*), which have received less theoretical and empirical attention than news stories or reports.

The study is based on the micro-thematic and the macro-thematic analysis of a bilingual sample consisting of a total of thirty two texts (sixteen English and sixteen Spanish) divided into two equal groups, one corresponding to news reports and the other to commentaries. The micro-thematic analysis focused on the clausal level and examined the experiential elements selected as Thematic Heads, their grammatical realization and internal structure. Textual and interpersonal choices were also inspected as part of a multiple Theme. The macro-thematic analysis examined the thematic progression patterns as well as the distribution of themes in text stages.

The results of the comparative analysis revealed a number of interesting differences both at the micro-thematic and macro-thematic levels between both genres, and certain language-specific preferences. To illustrate, at the micro-thematic level Sayers in Verbal processes are the preferred type of experiential role as Thematic Head in news reports in both languages, and Carriers in Relational processes are the preferred selection in commentaries. These two selections are deliberate choices on the writer's part in each genre. In news reports the writer tends to attribute information to outside sources to give an impression of factuality and objectivity. In commentaries the writer strives to present his views as unattributed evaluations, as opinions based on facts. However, the selection of the Process element as the first experiential element in clause-initial position in the Spanish clause is a typological feature of Spanish, a pro-drop language where the lexical realization of the Subject is not obligatory.

In view of the similarities and statistically-significant differences observed between both genres and languages, it is suggested that the influence of genre seems to be the major factor affecting the thematic variation observed in the bilingual sample, while language-specific differences play a secondary role.

References

Alonso, I. (coord.) (1997) Different Approaches to Newspaper Opinion Discourse. *Special Issue of Revista Electrónica de Lingüística Aplicada.* Asociación Española de Lingüística Aplicada (AESLA).

Dafouz, E. (2008). The pragmatic role of textual and interpersonal metadiscourse markers in the construction and attainment of persuasion: A cross-linguistic study of newspaper discourse. *Journal of Pragmatics* 40 (2008): 95-113.

Lavid, J., Arús, J. and L. Moratón (2010). Comparison and translation: towards a combined methodology for contrastive corpus studies. *International Journal of English Studies*. Pp. 159-173.

Tirkkonen-Condit, S. (1996). Explicitness vs. explicitness of argumentation: and intercultural comparison. *Multilingua* 15 (3), 257–273.

Wang, W. (2008). Intertextual aspects of Chinese newspaper commentaries on the events of 9/11. *Discourse Studies* 10 (3): 361-381.

Youngspeak: Spanish *vale* and English *okay*

Anna-Brita Stenström
Bergen University

As has been emphasized in a number of publications, pragmatic/discourse markers play a crucial role in colloquial language overall, and in young people's language in particular. Until now, the use of pragmatic markers in a contrastive perspective has been devoted less attention. Exceptions are, for instance, Aijmer & Simon-Vandenbergen (2006, Stenström (2006, 2008, 2009), Jørgensen & Stenström (2009). This paper is devoted to the Spanish pragmatic marker *vale* and its nearest English equivalent *okay,* as they are used in *Corpus Oral de Lenguaje Adolescente de Madrid* (COLAm) and *The Bergen Corpus of London Teenage Language* (COLT). The study shows among other things that, although both markers are multifunctional, *okay* is a more versatile marker than *vale.*

A problem that will be touched upon from a sociolinguistic point of view – is the difficulty involved in comparing data from corpora that, by necessity, are not identical in terms of gender, age and socioeconomic background information.

References

Aijmer, Karin and Anne-Marie Simon Vandenbergen (eds.). 2006*. Pragmatic Markers in Contrast. Studies in Pragmatics 2*. Oxford: Elsevier.

Jørgensen, Annette M. and Anna-Brita Stenström. In press. Dos marcadores pragmáticos contrastados:en el lenguaje juvenil: el inglés *like* y el español *como.* To appear in *Revista Español Actual.*

Stenström, Anna-Brita. 2006. The Spanish discourse markers *o sea* and *pues* and their English correspondences. In K. Aijmer and A-M. Simon-Vandenbergen (eds.). 155-172.

Stenström, Anna-Brita. 2008. Algunos rasgos caraterísticos del habla de contacto en el lenguaje de adolescentes en Madrid. *Oralia*: 207-226.

Stenström, Anna-Brita. 2009. Pragmatic markers in contrast: Spanish *pues nada* and English *anyway.* In A-B. Stenström and A. M. Jørgensen (eds.): *Youngspeak in a Multilingual Perspective*.Amsterdam: Benjamins, 137-160.

An analysis of translational complexity in English-Norwegian parallel texts

Martha Thunes
University of Bergen

The paper presents some results and topics from Thunes (forthcoming), a study of translational complexity in selected parallel texts of English and Norwegian. The investigation includes narrative fiction and law texts. The fiction texts are part of the English-Norwegian Parallel Corpus (ENPC), and the law texts are publicly available. The texts include about 68,000 words, and cover comparable amounts of texts of each type.

The analysis is applied to manually extracted pairs of translationally corresponding strings, where the finite clause is the basic unit of translation. The string pairs are classified according to a hierarchy of four correspondence types, reflecting the complexity of the relation between source and target string.

In type 1, the least complex type, the corresponding strings are pragmatically, semantically, and syntactically equivalent, down to the level of the sequence of word forms. In type 2, source and target string are pragmatically and semantically equivalent, and equivalent with respect to syntactic functions, but there is at least one mismatch in the sequence of constituents or in the use of grammatical form words. Within type 3, source and target string are pragmatically and semantically equivalent, but there is at least one structural difference violating syntactic functional equivalence between the strings. In type 4, there is at least one linguistically non-predictable, semantic discrepancy between source and target string. I.e., type 4 covers correspondences where the translation cannot be predicted from the source expression together with information about source and target language and their interrelations. Hence, such cases are probably outside of the scope of automatic translation. The method is previously presented in Thunes (1998), and adapted versions of it are used by Azevedo (in progress), Silva (2008), and Tucunduva (2007).

The distribution of the four correspondence types within the data is a measure of the degree of translational complexity in the analysed parallel texts. Across all data, type 4 correspondences cover more than half of the analysed texts. The law text data exhibit a larger proportion of semantic equivalence between source and target strings than the fiction data do, reflecting the fact that law texts are norm-governed in a way that fiction is not. However, the results show that also within the fiction texts there are differences concerning how far the original content is preserved in the target. Moreover, cases of semantic specification and despecification are frequent among type 4 correspondences. To some extent this reflects fairly systematic links between finite and nonfinite constructions. As for specification, its frequency indicates the level of explicitation in the various translations.

Primary sources

Law texts:

Agreement on the European Economic Area. Articles 1–99. 1992. The Norwegian Royal Ministry of Foreign Affairs.

Avtale om det Europeiske Økonomiske Samarbeidsområde. Artikler 1–99. 1992. The Norwegian Royal Ministry of Foreign Affairs.

Lov om petroleumsvirksomhet. §§1–65. 1994. The Norwegian Petroleum Directorate.

Act relating to petroleum activities. Sections 1–65. 1994. The Norwegian Petroleum Directorate.

Fiction texts:

Brink, André. 1984. *The Wall of the Plague.* Extract from beginning: pp. 13–23. London: Faber and Faber.

Brink, André. 1984. *Pestens mur.* Extract from beginning: pp. 11–20. Translated by Per Malde. Oslo: H. Aschehoug & Co (W. Nygaard) AS.

Hansen, Erik Fosnes. 1990. *Salme ved reisens slutt.* Extract from beginning: pp. 15–28. Oslo: J. W. Cappelens Forlag AS.

Hansen, Erik Fosnes. 1996. *Psalm at Journey's End.* Extract from beginning: pp. 7–18. Translated by Joan Tate. New York: Farrar, Straus and Giroux.

Lessing, Doris. 1985. *The Good Terrorist.* Extract from beginning: pp. 5–15. London: Jonathan Cape.

Lessing, Doris. 1985. *Den gode terroristen.* Extract from beginning: pp. 5–15. Translated by Kia Halling. Oslo: Gyldendal Norsk Forlag AS.

Vik, Bjørg. 1979. *En håndfull lengsel.* Extract from beginning: pp. 9–23. Oslo: J. W. Cappelens Forlag AS.

Vik, Bjørg. 1983. *Out of Season and Other Stories.* Extract from beginning: pp. 1–13. Translated by David McDuff and Patrick Browne. London: Sinclair Browne.

Secondary sources

Azevedo, Flávia. In progress. Investigating the problem of codifying linguistic knowledge in two translations of Shakespeare's sonnets: a corpus-based study. Poster presentation at ICAME 32, Oslo, 1–5 June 2011.

Johansson, Stig and Signe Oksefjell (eds). 1998. Corpora and Cross-linguistic Research: Theory, Method, and Case Studies. Language and Computers: Studies in Practical Linguistics 24. Amsterdam and Atlanta, GA: Rodopi.

Silva, Norma Andrade da. 2008. *Análise da tradução do item lexical evidence para o português com base em um corpus jurídico.* Master's thesis. Federal University of Santa Catarina, Florianópolis.

Thunes, Martha. 1998. Classifying translational correspondences. In: Johansson and Oksefjell (eds), 1998, 25–50.

Thunes, Martha. Forthcoming. *Complexity in Translation. An English-Norwegian Study of Two Text Types.* Doctoral dissertation. University of Bergen.

Tucunduva, Camila de Andrade. 2007. *Translating completeness: a corpus-based approach.* Master's thesis. Federal University of Santa Catarina, Florianópolis.

Seeing the lexical profile of Swedish through multilingual corpora

Åke Viberg
Uppsala University

After a peak in interest in the 1950s and 1960s, contrastive studies for a rather long time did not attract very much interest. In recent years, there has been a renewed interest which is closely related to the adoption of a new methodology, corpus-based studies of parallel corpora. Stig Johansson played a central role in this development. His book *Seeing through multilingual corpora* (Johansson 2007) is a good introduction to corpus-based contrastive studies. Personally, I had been working with lexical typology when I was invited to a conference on the English Norwegian and English Swedish Parallel Corpora he was building together with Karin Aijmer and Bengt Altenberg (see Aijmer et al. 1996) and became interested in the possibilities offered by corpus-based studies.

My own studies represent an attempt to combine lexical typology and corpus-based contrastive analysis to characterize the verb lexicon of Swedish and to identify language-specific features as a contribution to the characterization of the typological lexical profile of Swedish (see Viberg 2006). General typology with a world-wide scope provides the basic framework but the major focus is a contrastive comparison of Swedish and a selection of genetically and/or areally relatively closely related languages based on corpus-data which make it possible to provide a relatively fine-grained semantic analysis. Data are obtained from a parallel corpus that is being compiled by the author and is referred to as the Multilingual Parallel Corpus (MPC). At present, it consists of extracts from 22 novels in Swedish with translations of all texts into English, German, French and Finnish. For some texts, translations also are included into other languages. The Swedish original texts comprise around 600,000 words. In addition, there are original texts from some languages other than Swedish (e.g. Finnish) with Swedish translations. Data are also obtained from the English Swedish Parallel Corpus (ESPC) prepared by Altenberg & Aijmer (2000), which contains original texts in English and Swedish together with their translations.

A number of studies have already been completed on several verbal semantic fields in Swedish, such as mental verbs (Viberg 2005), verbs of perception (Viberg 2008), and verbs of possession (Viberg 2010). These studies have been based on earlier, more restricted versions of the MPC corpus. In the presentation, extended versions of a selection of representative examples will be supplied.

References

Aijmer, Karin, Altenberg, Bengt & Johansson, Mats. Eds. 1996. *Languages in contrast. Papers from a Symposium on Text-based Cross-linguistic Studies* [Lund Studies in English 88] Lund: Lund University Press.

Altenberg, Bengt & Aijmer, Karin. 2000. The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies. In Christian Mair and Marianne Hundt (eds.), *Corpus linguistics and linguistic theory*, 15-33. Amsterdam & Atlanta: Rodopi.

Johansson, Stig. 2007. Seeing through multilingual corpora. On the use of corpora in contrastive studies. Amsterdam: Benjamins.

Viberg, Åke. 2005. The lexical typological profile of Swedish mental verbs. *Languages in Contrast 5:1*, 121-157.

--- 2006. Towards a lexical profile of the Swedish verb lexicon. In: Viberg, Å. Guest editor. The Typological Profile of Swedish. Thematic issue of *Sprachtypologie und Universalienforschung. Vol. 59:1*, 103-129.

--- 2008. Swedish verbs of perception from a typological and contrastive perspective. In: María de los Ángeles Gómez González, J. Lachlan Mackenzie and Elsa M. González- Álvarez (Editors) *Languages and Cultures in Contrast and Comparison.* Amsterdam: John Benjamins, 123-172.

---. 2010. Basic Verbs of Possession. In: Maarten Lemmens (ed.) Unison in multiplicity: Cognitive and typological perspectives on grammar and lexis. *CogniTextes 4*. http://cognitextes.revues.org/308.

Workshop 2: Do we still need language corpora?

Martin Wynne, Ylva Berglund Prytz
University of Oxford

Language corpora were originally developed as datasets for linguistic research, in a world where researchers rarely had access to machine-readable language data. Pioneers such as Stig Johansson provided an invaluable service and helped to create a new paradigm in linguistic research. Corpus linguistics subsequently developed sets of procedures and methodologies based on discrete, bounded datasets, created to represent certain types of language use, and studied as exemplars of that domain. The growth of the field and advances in technology mean that corpora have become bigger and more plentiful and various, with huge reference corpora for a vast range of languages and time periods, and numerous specialist corpora representing a wide range of language varieties.

Nowadays, the enormous wealth of digital language data at our fingertips brings the role of the corpus into question. Large-scale digitization projects are delivering the writings of the past to our desktops in ways that allow us to configure *ad hoc*, bespoke datasets to help address our research questions. Much current language data is ‘born digital’, often a form of computer-mediated communication, and is easily captured and shared. Books and newspapers are published in electronic form, and made available in large collections. Online tools allow us to search for texts and collect them in virtual corpora. The boundaries between the corpus and other *ad hoc* datasets are blurring. What is the case for the carefully crafted corpus today?

The session will be a formal debate, with two speakers for and two against the motion, questions from the floor, and a summing up by the speakers, ending with a vote by the audience. The motion will be:

> “Language corpora are no longer necessary for linguistic research.”

Participants in the ICAME conference are warmly encouraged to come along and participate in what promises to be an entertaining debate on the key question confronting corpus linguistics today.

Speakers for the motion
Silvia Bernadini, University of Bologna
Elena Tognini-Bonelli, University of Siena

Speakers against the motion
Gregory Garretson, Uppsala University
Janne Bondi Johannessen, University of Oslo

Workshop 3: From multigenre to register-specific historical corpora

Merja Kytö, Uppsala University
Irma Taavitsainen, University of Helsinki

The aim of the workshop is to give up-to-date information about new developments and current trends in the versatile field of historical corpora. Register- and genre-specific corpora are often created to answer specific research questions, but they can be used for other research tasks and combined for a more comprehensive picture. Assessments of linguistic features across such databases show interesting distributions and can cast new light on core issues of historical linguistics.

Three centuries of drama dialogue: Compiling the ESDD corpus

Linnéa Anglemark
Uppsala University

Historical spoken and speech-like interaction is an intriguing subject. One text type that is well suited for research on speech-like interaction is drama dialogue, which is usually written to be spoken. Studies of dialogue in historical drama can help the researcher pinpoint and understand some changes in spoken interaction over time. To enable us to conduct such studies, researchers at the Department of English and the Department of Scandinavian Languages at Uppsala University have compiled the *English and Swedish Drama Dialogue* (ESDD) corpus. The aims of the project are to compare linguistic features and dialogic interplay in speech as represented in drama dialogue in English and Swedish, and to develop tools that will enable researchers to make these comparisons with the help of annotated corpora. The project thus involves several linguistic fields of study, including not only corpus linguistics but also historical sociolinguistics, historical pragmatics, and stylistics, with the added dimension of a cross-linguistic investigation, in that comparisons of e.g. pragmatic features, politeness strategies, or the frequency of pragmatic markers can be made between the two languages from a diachronic perspective.

The ESDD corpus comprises approximately 300,000 words drawn from English-language and Swedish-language dramas from three periods of time: 1725–1750, 1825–1850 and 1925–1950. Each language component contains 15 texts (five texts per sub-period) of 10,000 words in length. In each text, every utterance has been tagged. The tagging indicates a number of characteristics of the speaker and the addressee, including age, gender, and social status for both parties. This enables researchers to use search programs for investigating dialogic patterns and interplay in speech, as represented in drama dialogue, and to compare linguistic features in the language of characters. This presentation will focus on the development of the English-language part of the ESDD corpus, and on the software tools used for the project.

Being specific about historical change: The influence of sub-register

Douglas Biber
Northern Arizona University

This paper argues that historical linguistic change is mediated by register differences at a much more specific level than has been previously recognized. The paper begins by reviewing previous research that establishes the importance of general register differences for descriptions of both synchronic and diachronic patterns of linguistic variation. Historical research is especially problematic, because corpus materials are more difficult to collect, making it more difficult to ensure that the ‘same’ register is being compared across historical periods. In fact, the present paper argues that seemingly minor differences in register can correspond to meaningful differences in historical development. Two specific case studies from 20th century historical change are presented. The first compares the patterns of change in a corpus of articles from *Time Magazine* to those found in a corpus of articles from the *New York Times*, showing how the differing readerships and purposes of magazines versus newspapers results in different historical-linguistic patterns of use. The second case study then compares the patterns of change in three corpora of academic articles: professional science research articles, professional non-science research articles, and popular science articles. The last of these register categories is represented by articles from the *Philosophical Transactions of the Royal Society* and the journal *Science*; these journals have changed in purpose and readership over the last century, making them quite different from professional research articles in recent years. In conclusion, these case studies are interpreted relative to current practice in historical studies, to argue for more rigor in the standards of comparison required for comparative research.

Studying register(s) in the Corpus of 19th-century Scottish Correspondence

Marina Dossena
Università degli Studi di Bergamo (I)

The aim of this presentation is to provide an overview of the most relevant issues in the study of register in a historical corpus of correspondence.

The polyphony of voices and the multifarious purposes witnessed in letter exchanges make correspondence an ideal field in which to investigate both (sub)genre and register variation, as encoders and receivers are continuously seen to adapt their communicative strategies. In the case of the Corpus of 19th-Century Scottish Correspondence (19CSC) we are dealing with a collection of letters written by men and women of varying ages and of varying levels of education, for different purposes. The letters may be divided into different groups, according to their main function, and though it might be easy to make a distinction between familiar letters and business letters, it is not infrequent to come across instances in which dividing lines become much more blurred. In business letters more private issues, such as family health, may be mentioned; or familiar letters may deal with business matters, such as inheritance, leases, or money transfers, a frequent topic in emigrants’ letters.

My presentation will focus on instances in which the coexistence of different subgenres in the same collection may be investigated in terms of register variation, particularly in relation to the expression of greater psychological distance and/or greater professional competence. While specific attention will be given to lexical choices and modality, other indicators of register variation, such as (de)personalization strategies, will also be taken into consideration.

The Corpus of English Religious Prose – multigenre and register-specific

Thomas Kohnen, Tanja Rütten
University of Cologne

The religious domain does not only provide some of the most relevant and prestigious writings from the earliest records of the English language until well into the eighteenth century; it is also characterised by a highly complex network that involves different constellations of participants and numerous genres that either prevail or "enter" and "leave" the discourse community at various stages throughout its recorded history. Some genres have existed relatively unchanged ever since Old English times (e.g. sermons, liturgical prayers). Other genres and subtypes of genres appear at various stages of the English language (e.g. private prayers in late Middle English, catechisms and various treatise-types in Early Modern English). On top of that, new publication forms are adopted (e.g. pamphlets and newspaper articles with religious issues), and highly literate as well as primarily oral genres exist side by side (e.g. exegetical commentaries and catechisms).

In our presentation, we will shortly comment on the complexity of religious writing in terms of discourse functions, participants and relevance of the various genres, and then discuss the ways in which we may capture this complexity in historical corpora (in particular, in the *Corpus of English Religious Prose*). We introduce a matrix that includes three spheres and three different sets of genres and comment on our idea of a *network corpus* that is going to be register-specific and multigenre at the same time.

Corpora of Early English Correspondence (CEEC400)

Arja Nurmi
Research Unit for Variation, Contacts and Change (VARIENG)
University of Helsinki

The *Corpus of Early English Correspondence* has developed from one 2.6-million-word corpus to a corpus family covering 400 years of English correspondence from the beginning of the fifteenth century to the end of the eighteenth. The CEEC400 contains 5.2 million words of personal correspondence written by Englishmen and Englishwomen from all the literate social ranks. The corpus has been designed to be as socially representative as possible, and variables taken into account in compilation included gender, social status, geographic origin and relationship between writer and recipient. While based on edited letter collections, the compilation team has strived to discover reliable editions presenting the language of the original manuscripts in as authentic a way as possible. The corpora have a sender and recipient database, which contains information about the social backgrounds of informants. Some of this is also coded into the corpora (e.g. relationship between writer and recipient). A part of the corpus has been linguistically annotated and released for the use of the wider scholarly community.

The research carried out using the corpus started with the basic question: is it possible to apply the methods of present-day sociolinguistics to historical data?

Once this question was answered in the affirmative, the scope of research questions could be widened. Within the compilation team approaches have ranged from stratificational and interactive sociolinguistics to socio-pragmatics, but the corpus has also provided suitable material for more linguistically focused studies by scholars around the world.

The published parts of CEEC400 include the *Corpus of Early English Correspondence Sampler* (CEECS) and the *Parsed Corpus of Early English Correspondence* (PCEEC). Work is under way to clear copyrights for the publication of parts of the Corpus of Early English Correspondence Extension (CEECE) and the Corpus of Early English Supplement (CEECSU), hopefully in the course of 2011.

References

Corpora of Early English Correspondence. Entry at Corpus Research Database (CoRD). http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html

Nevala, Minna and Arja Nurmi (forthcoming). "The Corpora of Early English Correspondence (CEEC400)." *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*, ed. by Ville Marttila, Anneli Meurman-Solin, Carla Suhr and Jukka Tyrkkö. (Studies in Variation, Contacts and Change in English). http://www.helsinki.fi/varieng/journal/index.html

Representativeness revisited: Compiling a corpus of nineteenth-century newspaper English

Erik Smitterberg
Uppsala University

As studies such as Hundt and Mair (1999) and Westin (2002) have shown, newspaper English was an important locus of language change during the twentieth century; the language of twentieth-century newspapers has therefore been studied in some detail. In contrast, less is known about developments in nineteenth-century newspaper English, in spite of the fact that the newspaper was an influential medium during the 1800s: Lee (1976: 18) argues that "[t]he press in the nineteenth century was the most important single medium of the communication of ideas". There is thus a need for corpora that include nineteenth-century press language. The Corpus of Nineteenth-century Newspaper English (CNNE) is intended to be a source of data that can be used as a basis for investigations of nineteenth-century British newspaper English.

The aim of this presentation is to discuss some of the challenges associated with compiling a corpus such as CNNE. I shall address three main areas where decisions are called for during the compilation process:

1. Genre development. As Leech (2007: 143) notes, genres are subject to change over time. The British press changed a great deal during the course of the nineteenth century; for instance, the repeal of the Stamp Duty in 1855 "made the penny daily paper an economic possibility" (Brown 1985: 4). Such developments raise questions regarding the extent to which an issue of a newspaper from the early and another from the late nineteenth century can be considered examples of the "same" publication.
2. Subgenre coverage. Newspaper English comprises a variety of subgenres, such as editorials and advertisements, and there are linguistic differences between these subgenres. Corpora of newspaper English that contain different subgenres, or different proportions of the same subgenres, may thus not be comparable.

3. The use of Optical Character Recognition (OCR) technology. As the number of digital newspaper archives covering the nineteenth century continues to grow, OCR is likely to become an increasingly important part of corpus compilation in this field. However, using OCR imposes several constraints on text selection.

As I will show in the presentation, decisions made by corpus compilers pertaining to all three areas have consequences for the representativeness of the resulting corpus of texts.

References

Brown, L. 1985. *Victorian News and Newspapers*. Oxford: Clarendon Press.

Hundt, M., and Mair, C. 1999. "'Agile' and 'Uptight' Genres: The Corpus-based Approach to Language Change in Progress". *International Journal of Corpus Linguistics* 4(2), 221–242.

Lee, A. J. 1976. The Origins of the Popular Press in England: 1855–1914. London: Croom Helm.

Leech, G. 2007. "New Resources, or Just Better Old Ones? The Holy Grail of Representativeness". In: Hundt, M., Nesselhauf, N., and Biewer, C. (eds.), *Corpus Linguistics and the Web*. Amsterdam and New York: Rodopi, 133–149.

Westin, I. 2002. *Language Change in English Newspaper Editorials*. Amsterdam and New York: Rodopi.

Representing a text domain with fuzzy edges: Introducing EMEMT

Jukka Tyrkkö
Research Unit for Variation, Contacts and Change (VARIENG)
University of Helsinki

The *Early Modern English Medical Texts* corpus (EMEMT) was released in December 2010. The release of the corpus was the culmination of several years of work by members of the Scientific thought-styles project and several collaborating scholars. The corpus is the second part in a diachronic series of three corpora collectively entitled the *Corpus of Early English Medical Writing* (CEEM).

The EMEMT corpus represents medical writing from 1500 to 1700, a period of two centuries during which the medical profession underwent fundamental changes on many different levels. In compiling the corpora, the team emphasized the importance of comprehensive understanding of relevant bibliographic and biographic metadata. This dedication to the philological approach is evident in the wealth of background information provided with the corpus. Although representative in the first instance of medical writing, EMEMT is broad enough in scope to allow for studies of Early Modern scientific writing in general.

EMEMT was published on CD-ROM with an accompanying book (eds. Taavitsainen and Pahta) giving details about the structure of the corpus, its sociocultural background, and the corpus tool. The book includes chapters by all members of the project team.

References

Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr and Jukka Tyrkkö (compilers), with the assistance of Alpo Honkapohja, Anu Lehto and Raisa Oinonen. 2010. *Early Modern English Medical Texts (EMEMT, 1500–1700).* CD-ROM with EMEMT Presenter software by Raymond Hickey. Amsterdam and Philadelphia: John Benjamins.

Taavitsainen, Irma and Päivi Pahta (eds.). 2010. *Early Modern English Medical Texts: Corpus Description and Studies*. Amsterdam and Philadelphia: John Benjamins.

Full papers
Software presentations
Work-in-progress reports

Pragmatic variation across corpora: the use of general extenders in different varieties of English

Karin Aijmer
University of Gothenburg

Pragmatic markers can be described as a 'fragile area of the linguistic system' (Mair 2006). They can for instance vary with regard to such factors as region and social factors such gender, social class and age. The present study takes its inspiration from variational pragmatics (Schneider and Barron 2008), which has the goal of examining pragmatic variation across geographical and social varieties. New pragmatic markers or new uses of pragmatic markers in one variety of English travel quickly to other regional varieties. This will be illustrated with 'general extenders' such as *and stuff* (*like that*), *and all (that)*, *or something* and their variants (cf. also Pichler and Levey fc).

The general extenders will be compared on the basis of their frequencies in the International Corpus of English (the corpora for British, American English, Australian English, Canadian English). Although all the varieties have extenders which can be analysed in the same structural frames, there are both formal and functional differences in their use which will be discussed in terms of regional and social factors.

References

Mair, Christian. 2006. Corpus linguistics meets sociolinguistics: the role of corpus evidence in the study of sociolinguistic variation and change. In: Renouf, A. and A. Kehoe (eds), Corpus Linguistics: Refinements and Reassessments. Amsterdam and New York: Rodopi.

Overstreet, Maryann. 1999. *Whale, candlelight and stuff like that: General extenders in English discourse.* Oxford: Oxford University Press.

Pichler, Heike and Stephen Levey. Forthcoming. A variationist perspective on the grammaticalization of general extenders in north-east England.

Schneider, Klaus P. and Anne Barron (eds.) 2008. *Variational pragmatics: A focus on regional varieties in pluricentric languages.* Amsterdam & Philadelphia: John Benjamins, 2008.

Rhetorical tag questions

Karin Axelsson
University of Gothenburg

Tag questions have traditionally been described as inviting verification or confirmation (Quirk *et al.* 1985:811). However, in a study of tag questions in spoken conversation and fiction dialogue in the *British National Corpus*, most tag questions turn out not to elicit responses, i.e. they are predominantly used rhetorically: three quarters of tag questions in spoken conversation and two thirds of tag questions in fiction dialogue have been analysed as rhetorical.

Rhetorical questions are in general "*special uses of questions*, rather than separate types of questions" (Ilie 1994:77); hence, the formal features of rhetorical tag questions are similar to response-eliciting tag questions, the main difference instead being that the speaker often continues speaking after a rhetorical tag question, particularly in fiction dialogue. Moreover, rhetorical tag questions are commonly accompanied by pragmatic markers such as *I mean*, *you know*, *I think*, *I* *suppose*, *well*, *really*, *though*, *anyway* and *actually* (cf. Brinton 1996:32), particularly in spoken conversation.

Rhetorical tag questions may be divided into *speaker-centred* tag questions, where the convictions and assessments of the speaker are in focus, and *addressee-oriented* tag questions, where the addressee is crucial: the speaker then often presents suppositions concerning the addressee, but the addressee may also, for example, be reminded, accused or challenged. Most rhetorical tag questions are speaker-centred in spoken conversation but addressee-oriented in fiction dialogue. Speaker-centred rhetorical tag questions in spoken conversation deal quite often with uncontroversial everyday matters where the speaker assumes the addressee to have similar views, the tag sometimes seeming like a routine-like addition, whereas speaker-centred rhetorical tag questions in fiction dialogue mostly present controversial stance-taking on important matters as part of an argumentation to confront or convince the addressee. The predominance of addressee-oriented rhetorical tag questions in fiction dialogue appears to reflect an interest in depicting personal relationships, and in particular, conflicts and confrontations between characters: hence, addressee-oriented rhetorical tag questions challenging and accusing the addressee are more common in fiction dialogue than in conversation.

References

Brinton, Laurel. 1996. *Pragmatic markers in English.* Berlin: Mouton de Gruyter.

Ilie, Cornelia. 1994. *What else can I tell you? A pragmatic study of English rhetorical questions as discursive and argumentative acts*. Stockholm: Almqvist & Wiksell International.

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

Did *go/come*-V oust *go/come-and*-V? A study of the diachronic development of both constructions in American English

Ingo Bachmann
University of Duisburg-Essen

In their base form, the verbs *go* and *come* allow two different complementation patterns, as illustrated by the following sentences:

But I say, Paul, ***go and get*** *your hat, and we'll go out for a walk.* (COHA, 1865)
Oh Bunker, ***come and get*** *us! cried Sue.* (COHA, 1920)
How about I ***go get*** *you some food?* (COHA, 1996)
I'll just phone Sabrina and ask her to ***come get*** *us.* (COHA, 1999)

When it comes to the status of *go/come-and*-V and *go/come*-V in American English, we find that contemporary spoken American English privileges *go/come*-V over *go/come-and*-V (Biber et al. 1999: 1031, Mittmann 2004: 167).

But this synchronic snapshot misses the fact that this has not always been the case. The present empirical study of the diachronic development of *go/come-and*-V and *go/come*-V in 19th and 20th American English texts, taking the *Corpus of Historical American English* (compiled by Mark Davies) as its data basis, shows that both constructions underwent a remarkably diverging development. Whereas *go*-V only started to rise significantly in frequency at the turn of the 20th century, displaying a more or less steady increase up to today's norm, *go-and*-V dropped in frequency after having its peak in the second half of the 19th century. The same tendency can be observed for *come*. The question is: Is this just chance or did *go/come*-V oust *go/come-and*-V? And if they did so, why?

To provide answers to these questions, the following issues will be tackled:

- Do both constructions favour the same grammatical context? Are both constructions distributed similarly over the registers available in the corpus? Are they used with the same set of verbs? In summary, are they exchangeable or has each construction acquired a specialised function over time?
- Focusing on the rising use of *go/come*-V, which contexts took the lead? Can we see an increase first with today's most frequent verbs, such as *get* and *see*? Are there different developments depending on the verb form of *go* and *come*?

References

Biber, D. et al. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Mittmann, B. 2004. *Mehrwort-Cluster in der englischen Alltagskonversation*. Tübingen: Narr.

Using verifiable author data: Gender and spelling differences in Twitter and SMS

Alistair Baron (Lancaster University), Caroline Tagg (Open University), Paul Rayson (Lancaster University), Phil Greenwood (Lancaster University), James Walkerdine (Lancaster University), Awais Rashid (Lancaster University)

The web is increasingly being used as a corpus source for a wide range of linguistic studies (see Kilgarriff and Grefenstette, 2003; Hundt et al, 2007). One area of the web which has received particular recent interest is Twitter, a micro-blogging site where users can post messages, or “tweets”, about any subject; including personal news updates, comments, responses to other messages, adverts and continuing online conversations. The site has 175 million users with 95 million tweets written each day (as of September 14th 2010[1]) and, with the data being publicly available, it is easy to see why Twitter is becoming popular as a resource for corpus linguistics.

One issue with using web-based data in a corpus is the general inability to attach reliable author metadata to specific texts. Researchers rely on users entering their details accurately and honestly (e.g. Schler et al, 2006), but many studies have found that Internet users often mask their true identity for a variety of reasons (e.g. Danet, 1998; Donath, 1999; Gross, 2004). We have attempted to avoid this issue by compiling a Twitter corpus containing only “verified users”, that is users that Twitter staff have checked are actually who they say they are. Because the majority of these users are “famous” in some way, we were able to manually attach metadata by searching for the relevant information for each person online. The resulting 10 million word corpus allows us to study language differences on Twitter based on a user’s gender, age and English variety with relatively high confidence that the attached metadata is accurate.

Here, as well as describing the Twitter corpus and how researchers may compile their own similar corpus, we also present a case study using the Twitter corpus and an SMS corpus (Tagg, 2009), which also contains author metadata, to investigate gender dependent language differences, particularly in terms of orthography used. It has been shown previously how two tools, VARD (VARiant

Detector) and DICER (Discovery and Investigation of Character Edit Rules), can be used to investigate spelling trends in SMS (Tagg et al, 2010). Here, we extend this investigation to Twitter but also look closely at the differences between how spellings and other orthographic features, such as emoticons and punctuation, are used by the two genders in both corpora.

[1] Taken from http://twitter.com/about.

References

Danet, B. (1998). Text as mask: gender, play, and performance on the internet. In Jones, S. G. (ed.), *Cybersociety 2.0: revisiting computer-mediated communication and community*, chapter 5, 129–158. Sage Publications, Thousand Oaks, CA, USA.

Donath, J. (1999). Identity and deception in the virtual community. In Kollock, P. and M. Smith (eds.), *Communities in Cyberspace*, 29–59. Routledge, London.

Gross, E. F. (2004). Adolescent internet use: What we expect, what teens report. *Journal of Applied Developmental Psychology*, 25(6), 633 – 649.

Hundt, M., Nesselhauf, N. & Biewer, C. (eds.). (2007). *Corpus linguistics and the web*. Amsterdam: Rodopi.

Kilgarriff, A., Grefenstette, G. (2003). Introduction to the special issue on the Web as Corpus. *Computational Linguistics*, 29(3), 333–347.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006). Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Tagg, C. (2009) *A Corpus Linguistics study of SMS text messaging*. PhD thesis, University of Birmingham.

Tagg, C., Baron, A. and Rayson, P. (2010). "I didn't spel that wrong did i. Oops": Analysis and standardisation of SMS spelling variation. In ICAME 31 Abstracts, 108–109, Gießen, Germany.

Verb complementation in South Asian English(es): the range and frequency of "new" ditransitives

Tobias Bernaisch, Christopher Koch
Justus Liebig University Giessen

The extent to which South Asian varieties of English are marked by a shared pool of structural characteristics is still an unresolved issue. While some scholars (cf. e.g. McArthur 2002) advocate a uniform stock of pan-South Asian features, others (cf. e.g. Gargesh 2009) stress the local character of the varieties of English used in the South Asian region. Likely reasons for this sustained disparity are a) the strong focus on the systematic description of Indian English, the largest second-language variety of English world-wide, at the cost or complete neglect of other established South Asian varieties, and b) the lack of empirically sound and sufficiently large authentic databases of the respective varieties.

The present paper reports on the design of a recently compiled 18-million-word newspaper corpus comprising acrolectal English language data from Bangladesh, India, the Maldives, Nepal, Pakistan and Sri Lanka and investigates verb complementation in the six varieties of English. Building on earlier research by Olavarría de Ersson & Shaw (2003), Nihalani et al. (2004) and Mukherjee & Hoffmann (2006), we examine "new" ditransitives, i.e. verbs which are (claimed to be) attestable in the ditransitive construction (cf. Goldberg 2005) in the South Asian varieties at hand (e.g. *he informed her the time*, *I put him a question*), but not in the present-day version of their historical input variety British English. As the range and frequency of "new" ditransitives can be regarded as markers of structural nativisation (cf. Schneider 2003, 2007), the analysis offers insights into norm developments in the South Asian region.

The study provides the first systematic analysis of "new" ditransitives across various South Asian Englishes. The overall picture is complex: while some verbs are used in the ditransitive construction in a number of South Asian varieties (pan-South Asian features), others only occur in individual varieties (variety-specific features). Apart from a qualitative inspection of these verbs and their corresponding structures, we will also report on quantitative differences between the six varieties with regard to the frequencies (i.e. the degree of pervasion) of these "new" ditransitives.

References

Gargesh, R. (2009): "South Asian Englishes". *The Handbook of World Englishes*, eds. B.B. Kachru, Y. Kachru & C. Nelson. Malden: Wiley-Blackwell. 90-113.

Goldberg, A.E. (1995): *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: The University of Chicago Press.

McArthur, T. (2002): *The Oxford Guide to World English*. Oxford: Oxford University Press.

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.

Nihalani, P., R.K. Tongue, P. Hosali & J. Crowther (2004): *Indian and British English: A Handbook of Usage and Pronunciation*. 2nd ed. Delhi: Oxford University Press.

Olavarría de Ersson, E. & P. Shaw (2003): "Verb complementation patterns in Indian Standard English". *English World-Wide* 24: 137–161.

Schneider, E.W. (2003): "The dynamics of New Englishes: from identity construction to dialect birth". *Language* 79(2): 233-281.

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

Investigating "collocational richness" in translated and non-translated language

Silvia Bernardini
University of Bologna

This paper describes a method for comparing originals and translations in the same language in terms of the number of collocations they contain. Viewing translation as being of relevance mainly to the target language (in line with mainstream research in translation studies, Toury 1995), adopting corpus-driven and corpus-based methods, and relying on monolingual comparable and bilingual parallel corpora of English and Italian, the paper aims to answer two research questions: Are translated texts richer/poorer in collocations than original texts in the same language? If so, is this a consequence of the translation process?

To answer the first question, all bigrams falling into predefined Part-Of-Speech patterns (e.g. Adjective - Noun sequences), and regardless of their frequency, are extracted from a sub-corpus containing translations and from its comparable counterpart containing non-translated texts in the same language. They are then ranked based both on their joint frequency and on their Mutual Information as observed in a reference corpus, and the rankings are compared.

Where a significant difference is found – suggesting that originals contain more / stronger collocations matching that POS pattern than translations or vice versa – bilingual concordance data are perused. This parallel phase of the research aims to answer the second question, i.e. to ensure that observed differences are motivated by the translation process (rather than unrelated variables), and to shed some light, albeit indirectly, on the underlying decision-making processes.

This method is applied to three sets of very small corpora: a bidirectional corpus of original and translated fiction extracts in English and Italian following the design of the ENPC (Johansson 2007), a corpus of original software documentation texts in original English, translated Italian and original Italian, and a corpus of shareholder letters from annual reports of multinationals in original Italian, translated English and original English. Results suggest that translated fiction displays a preference for collocations with respect to its comparable original counterpart, and this tendency is stronger for Italian than it is for English. The non-fiction corpora provide a less clear-cut picture.

While the parallel phase of the research is clearly more specific to research on translation, the simple monolingual method employed to evaluate collocational richness described here can be applied to any monolingual corpus comparison and even to the comparison of single texts, and is especially valuable for overcoming the well-known data sparseness issue afflicting collocational studies based on small amounts of textual data.

References

Johansson, S. 2007. *Seeing Through Multilingual Corpora*. Amsterdam: Benjamins.

Toury, G. 1995. Descriptive Translation Studies and Beyond. Amsterdam: Benjamins.

Talking about the past in Cook Islands English

Carolin Biewer
University of Zurich

Cook Islands English (CookE) is a second-language variety of English spoken in the Cook Islands in the South Pacific. One prominent feature in which this variety differs from Standard British English is the occasional lack of past tense marking on the verb, as in *Last year I learn at school...*. Several factors may influence the preference for such an omission:

(a) In Maori – the native language of the Cook Islanders – tense is only signalled by free morphemes in form of a preverbial particle (Lynch 1998: 130f, Carpentier & Beaumont 1995: 29ff). At the same time speakers of English as a second or foreign language usually reduce structures of the target language at the beginning of the learning process, in particular bound morphemes, and tend to avoid redundancies (Winford 2003: 218, Williams 1987: 174f). This may result in the preference of a non-marking on the verb if a time adverbial as a free morpheme is present.

(b) In the Oceanic substrate a particle to signal tense is not a compulsory sentence element if tense has already been marked earlier on during the conversation (Lynch 1998: 133). This particularly applies to informal narratives in the Oceanic substrate (as found by Dixon (1988: 69) for Fijian). This may result in the preference of a non-marking on the verb when the speaker tells a story.

(c) Phonological factors can also be of influence here as consonant clusters tend to be avoided in CookE, as the substrate does not have consonant clusters (Lynch 1998: 83).

(d) Furthermore, it would be interesting to see whether social variables such as gender, age and regional upbringing have an effect on the frequency of non-marked verbs in past tense contexts.

This paper will discuss the various internal and external factors that may trigger variation in the use of past tense marking in CookE, such as preceding/following phonological segment, regular/irregular verb, presence/absence of time adverbial, part/not part of a narrative sequence, gender, age and regional upbringing. The VARBRUL program will be used to perform a variable rule analysis to assess the strength of these possible constraints and their combined effects. Data will be taken from recordings made on a field trip to the Cook Islands in 2007. Although (t, d) deletion is a well-researched variable in variationist sociolinguistics (e.g. Guy 1977, Tagliamonte 2005), it is worthwhile to add another study on a non-native variety of English – not only because no research has been done on this variety apart from Biewer (2008, 2009) but also because such as study will give us new insights into a possible ESL-ENL distinction in the way we talk about the past.

References

Biewer, Carolin 2008. "South Pacific Englishes – unity and diversity in the usage of the present perfect." In Terttu Nevalainen et al. (eds.): *Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*. Amsterdam: John Benjamins, 203-219.

---- 2008. "Concord patterns in South Pacific Englishes – the influence of New Zealand English and the local substrate." In Stierstorfer, K. (ed.): *Anglistentag 2007 Münster, Proceedings*. Trier: Wissenschaftlicher Verlag, 331-343.

---- 2009. "Modals and semi-modals of obligation and necessity in South Pacific Englishes." *Anglistik* 20 (2): 41-55.

Carpentier, Tai T. T. & Clive Beaumont 1995. *Kai Koorero – A Cook Islands Maori Language Coursebook*. Auckland: Pasifika Press.

Dixon, R. M. W. 1988. *A Grammar of Boumaa Fijian.* Chicago: University of Chicago Press.

Guy, Gregory. 1977. "A new look at –t, –d deletion." In R.W. Fasold & R.W. Shuy (eds.): *Studies in Language Variation.* Washington DC: Georgetown University Press: 1-11.

Lynch, John 1998. *Pacific Languages. An Introduction.* Honolulu: Hawaii.

Tagliamonte, Sali. 2005. "New perspectives on an ol' variable: (t,d) in British English." *Language Variation and Change* 17: 281-302.

Williams, Jessica 1987. "Non-native varieties of English: A special case of language acquisition." *English World-Wide* 8 (2): 161-199.

Winford, Donald 2003. *Introduction to Contact Linguistics*. Oxford: Blackwell Publishing.

The competition between the intensifiers *dead* and *deadly*: synchronic and diachronic considerations

Zeltia Blanco-Suárez
University of Santiago de Compostela

One of the most susceptible areas to linguistic renewal and change is that of intensification, and within this domain in particular, the so-called intensifiers (cf., among others, Bolinger 1972, and Tagliamonte 2008). Being linguistic fads, therefore, intensifiers constitute a highly productive topic in scholarly discussion (cf., for instance, Paradis 1997; Macaulay 2002; Buchstaller and Traugott 2006, and Méndez-Naya 2008).

This paper also tackles the intensifier strategy. By adopting a grammaticalization approach, it focuses on the evolution of two death-related intensifiers, namely the *-ly* form *deadly* and its zero-adverb counterpart *dead*, both meaning 'utterly', 'extremely', as in (1) and (2):

(1) *I'm **deadly** serious, this isn't a game*! (*LDOCE*, s.v. *deadly* adv.).
(2) *He was **dead** good-looking*. (*LDOCE*, s.v. *dead* adv.).

The first recorded examples of these two forms in the *OED* date from the 14th and the 16th centuries respectively; cf. (3) and (4):

(3) *I þat es sa **dedli** dill* ['stupid']. (a. 1300. *OED*, s.v. *deadly* adv. 4).
(4) *Oh he is olde dogge at expounding, and **deade** sure at a Catechisme*. (1589. *OED*, s.v. *dead*, adv. 2a).

This paper provides a detailed account of how these forms evolved over time, from original descriptive meanings (as in (5) and (6)), to gradually more subjective meanings (7) and (8)), finally becoming grammaticalized as intensifiers (cf. examples (1)-(4) above). In order to trace the diachronic evolution of the two competing forms, I resort to a variety of historical corpora, including the *HC* and the *CLMETEV*.

(5) *He was alle assmayhydde* ['upset']..& *fell doune to þe grounde **ded** asswo* ['unconscious']. (a.1450. *MED*, s.v. *ded* adv.).
(6) *He wonded þe kyng **dedely** fulle sore*. (a.1400. *MED*, s.v. *dedli*, adv.).
(7) *As **dead**-still as a marble man*. (1818. *OED*, s.v. *dead* adv).
(8) *Custaunce, with a **dedly** pale face..toward hir ship she wente*. (c.1390. *MED*, s.v. *dedli* adv.).

From a synchronic perspective, the paper also analyses the variation between the two forms in Present-Day English, examining the collocations in which these adverbs occur. Using data from the *BNC* and the *COCA*, attention is drawn here to the potential existence of certain structural, semantic, and/or pragmatic factors predisposing the choice of one adverb over the other in the contemporary language. The results obtained can thus complement recent studies (cf. Macaulay 2006; Barnfield and Buchstaller 2010), which show that *dead* rose exponentially in the 1990s and then dropped significantly at the turn of the century.

References

Barnfield, Kate, and Isabelle Buchstaller. 2010. 'Intensifiers on Tyneside: longitudinal developments and new trends'. *English World-Wide* 31 (3): 252-287.

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

Buchstaller, Isabelle, and Elizabeth C. Traugott. 2006. '*The lady was al demonyak*: historical aspects of Adverb *all*'. *English Language and Linguistics* 10 (2): 345-370.

Macaulay, Ronald. 2002. 'Extremely interesting, very interesting, or only quite interesting?' Adverbs and social class. *Journal of Sociolinguistics* 6 (3): 398-417.

Macaulay, Ronald. 2006. 'Pure grammaticalization: the development of a teenage intensifier'. *Language Variation and Change* 18 (3): 267-283.

Méndez-Naya, Belén. 2008. 'On the history of *downright*'. *English Language and Linguistics* 12 (2): 267-287.

Paradis, Carita. 1997. *Degree modifiers of adjectives in spoken British English.* Lund: Lund University Press.

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

Sources

BNC = *British National Corpus*. Aston, Guy, and Lou Burnard. 1997. *The BNC handbook*: *Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

COCA = *Corpus of Contemporary American English*. Davies, Mark. 2008. Available at: http://www.americancorpus.org.

LDOCE = *Longman Dictionary of Contemporary English*. 2003. Harlow: Pearson Education Limited.

MED = Kurath, Hans et al. (eds.). 1952-2001. *Middle English Dictionary*. Ann Arbor, MI: University of Michigan Press. Online version. Available at: http:// *quod.lib.umich.edu/m/med/*.

OED = *Oxford English Dictionary*. (1989). (2nd ed.). Oxford: Oxford University Press. Online version with revisions. Available at: http://www.oed.com.

Discourse markers of reformulation – A phraseological approach to *that is to say* in present-day English

Melanie Borchers
University of Duisburg-Essen

It has been commonly agreed that reformulators such as *that is to say*, *that is* and *i.e.* represent a subgroup of discourse makers (e.g. Blakemore 2007). They announce reinterpretations of the message conveyed in the previous discourse segment. They thus function as linkers and might therefore be considered coherence features.

According to the discourse function(s) they fulfil, they can be categorised into four major classes of reformulation: expansion, modification, reassessment and compression (cf. e.g. Milagros Del Saz Rubio 2007: 17). While some of these reformulators can clearly be grouped together and only show features common to one of these classes, others, like *that is to say*, prove to be multi-functional markers within their category.

While the use of *that is to say* shows a strong tendency of decline in American English (cf. data from the *Corpus of Contemporary American English* as well as from the *Corpus of Historical American English*), the *Lancaster-Oslo/Bergen Corpus* and the *British National Corpus* provide evidence in favour of its rather constant recent use in the other major variety of English.

By approaching the reformulator *that is to say* in these corpora from a phraseological point of view, new pieces of information are gathered to shed more light on the use and function of multi-word discourse markers.

About this time, too — that is to say, between 1976 and 1978 — it was established on impeccable authority to everyone's satisfaction that the Tibbu were Jews. (*BNC* ADW 193)

An older book, that is one published before around 1900, will only have black and white plates, which are unlikely to be photographs. (*BNC* A04 710)

The only Roman coins to bear an explicit date were some very rare ones made by the Emperor Hadrian in 'the year of the city 874' (i.e., AD121) and some made by the usurper Pacatian ‘in the 1001st year of Eternal Rome’ (i.e., AD249). (*BNC* ADH 228)

Phraseological principles, as for example opaqueness of meaning, that is the degrees of idiomaticity, or the degree of frozenness (cf. Fraser 1970) of the reformulator under investigation, enable us to answer the question, whether *that is to say*, *that is* and *i.e.* represent different variants of the same marker in the three preceding examples or whether they should be considered different discourse markers after all. In how far is it possible to add to or alter the reformulator (e.g. **that will be to say* or ?*that’s to say*) in order to keep the meaning constant?

References

Blakemore, Diane. 2007. “‘Or’-parentheticals, ‘that is ‘-parentheticals and the pragmatics of reformulation.” *Journal of Linguistics* 43: 311-339.

Fraser, Bruce. 1970. “Idioms within a Transformational Grammar.” *Foundations of Language* 6: 22-42.

Milagros Del Saz Rubio, Maria. 2007. *English Discourse Markers of Reformulation.* Bern: Peter Lang.

***From the very beginning*: the development of *very* in definite noun phrases**

Tine Breban
University of Leuven

English *very* is known to have a variety of modifying uses, including its use as degree modifier of adjectives and quantifiers, e.g. *very pretty girls*, *very many children*, and less frequent uses as emphasizer, e.g. *the very end*, and secondary determiner, *the very man I wanted to see* (see e.g. Brugman 1989, Breban, Davidse & Vandewinkel 2007, González-Diáz 2005). Historical corpus research (Adamson & González-Diáz 2005) has added two uses as gradable attributive adjective: an objective use “real, corresponding to fact” and a subjective one “true, genuine”. It was proposed that the degree modifier and emphasizer uses are the result of processes of grammaticalization and subjectification. Two specific paths of change were suggested: (1) objective attribute > subjective attribute > degree modifier, and (2) objective attribute > subjective attribute > emphasizer. How the secondary determiner use of *very* fits in with these developments has not yet been discussed. It is this question that I address on the basis of a new set of data from the Penn-Helsinki Parsed Corpora of ME and Early MoE, which is tailored towards the earliest secondary determiner uses. The data set consists of all attestations of *very* in definite noun phrases in the period 1420-1710 in which the first examples of its secondary determiner use were attested. The method to reconstruct possible paths of grammaticalization is to look for bridging contexts (Evans & Wilkins 2000), i.e. examples in which two readings are equally plausible and which illustrate how two meanings could have been related, and “onset contexts” (Traugott forthcoming). The data show that the secondary determiner use encompasses two subtypes, an emphasizing use, in which *very* highlights the fact that the identification of the referent is correct, e.g. *that is the very note* “indeed the note” *of it*, and a referential use, in which *very* sets up a link of co-referentiality with an antecedent referent (equivalent to PDE *same*), e.g. *thou lovedest me before the makynge of the worlde. [...], the very worlde hath not knowen the.* In terms of their relative chronology, the emphasizing secondary determiners appear to be related through bridging contexts with several objective, subjective attribute and emphasizer uses and later themselves constituted the input for the development of the referential use. Onset contexts for this shift are examples in which *very* emphasizes another marker of co-reference such as *same* or a demonstrative, e.g. *I falsely accused by the very same accusers*.

Corpus development, Lexis and the German Learner – Do nine-year olds talk about making their beds?

Martina Bredenbröcker, Ilka Mindt
University of Potsdam

Teaching standards for the EFL classroom are given and controlled by governmental departments. In how far should these teaching standards take into account results based on empirical research?

This work-in-progress report focuses on English language standards as prescribed by state authorities in Germany and tries to determine the adequacy of this material by using methods from corpus linguistics.

The Bavarian syllabus for teaching English at primary school level will be taken as an example. A sample of verbs and nouns was selected from this syllabus to investigate their collocations. This analysis was based on the British National Corpus (BNC). In a second step, the results from this corpus-based analysis were compared with the collocations given in the Bavarian syllabus.

Collocations have been a long-standing issue in corpus linguistics and EFL such as the analysis of collocations for learner dictionaries (Sinclair, 1991), their investigation in learner corpora (Altenberg/Granger, 2001 and Nesselhauf, 2005), and the study of frequency information of collocates in American English (Davies/Gardner 2010).

The collocations given for the verb *make* in the Bavarian syllabus are *homework* and *bed*. Corpus-based results reveal that *homework* does not occur together with the verb *make* at all. The noun *bed* is used more frequently with verbs such as *lie* or *sleep* than with *make*.

A critical discussion follows that takes into account frequency information, (proto-)typicality of use and linguistic considerations important for the classroom. After this, a research project will be sketched which involves compiling a corpus that reflects the lexis used by eight to ten year old English children. This corpus, together with standard corpora, will then be the basis for an investigation of lexical structures relevant for (German) learners of English.

References

Altenberg Bengt/Granger Sylviane (2001). "*The grammatical and lexical patterning of make in native and non-native student writing.*" *Applied Linguistics*, *22* (2), 173-194.

Amtsblatt der Bayerischen Staatsministerien für Unterricht und Kultus und Wissenschaft, Forschung und Kunst. Teil 1.Sondernummer 1 (KWMBl. I So.-Nr. 1/2000). München.

Davies, Mark/Gardner, Dee (2010). *A Frequency dictionary of contemporary American English*. New York: Routledge.

Hunston, Susan (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Nesselhauf Nadja (2005). *Collocations in a learner corpus*. Amsterdam & Philadelphia: Benjamins.

Sinclair John (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

http://www.isb.bayern.de/isb/download.asp?DownloadFileID= (22.11.10)

There's talk that a new construction of hearsay evidentiality is emerging in English

Lieselotte Brems, Kristin Davidse, An van Linden
University of Leuven

Among the Germanic languages, English stands out as lacking a modal auxiliary that expresses hearsay evidentiality (cf. Chafe 1986: 268–269; Brône and Feyaerts 2002).

In German, for instance, *sollen* ('shall, to be to') and *wollen* ('want') can be used to indicate that the speaker is not the original source of the reported proposition (cf. Mortelmans 2000). Present-day English, however, has to take recourse to other expression types, which have mainly lexical origins. Previous studies have reported on constructions like *be said/reported/alleged to* (Noël 2001), perception verb constructions with *hear* (Whitt 2009), adverbs like *allegedly* and *purportedly* (Chafe 1986), and hearsay uses of *seem to* (Aijmer 2009). In this paper, we argue that yet another construction of hearsay evidentiality is emerging, realized by the highly productive macro-construction '*there*+BE+'evidence'-noun+ complement clause' (whose use for the expression of modal meaning, as in e.g. *there is no doubt*, is discussed by Simon-Vandenbergen (2007)):

(1) *The sergeant refuses to elaborate but there is talk of crutch prints being left at the scenes of the crimes* (WO)

(2) *There were unconfirmed reports that on her otherwise slender body she owned two marvelously full breasts* (WO)

Based on data from *Wordbanks Online* (WO), this paper will present a first breadth study charting the synchronic variation attested in this emerging pattern both qualitatively and quantitatively. We will investigate the set of nouns used (as well as their premodification and determination) and the types of complement clauses (finite or non-finite, modalized or non-modalized). It will be argued that at present '*there*+BE+'evidence'-noun+complement clause' qualifies as a partially filled construction (cf. Goldberg 2006), in which some slots, such as impersonal *there*, are fixed, while others are (semi-)open, e.g. *be* can be conjugated, nouns can be *talk*, *rumour(s)*, *report(s)* or more infrequent nouns such as *rumblings* or *whispers*. We will also describe emergent phraseologies, such as *unconfirmed reports*, which allow for more fine-grained qualifications of evidentiality.

In addition, we will present a diachronic case study, based on the Helsinki-corpus and the Corpus of Late Modern English, in which we will argue that this pattern is the result of (ongoing) grammaticalization processes interacting with lexicalization (cf. Lehmann 2004). We will trace what appear to have been the earliest instantiations of this pattern and the specific semantic-pragmatic environments in which the evidential meaning emerged. We will also investigate the analogy relation between models and more or less frequent constructs attracted by them (cf. Hoffmann 2004). Observable changes in premodification and determination patterns of the 'evidence' nouns, and their specialization towards uncount and plural nouns, will be linked to decategorialization reflexes of constructionalization.

Patterns of wording to patterns of meaning

Mark Brown
Norwegian School of Management (BI)

In my paper I present a corpus-based attempt to identify and compare patterns of wording in the lexicogrammar of two distinct discourse communities. This empirical exercise was inspired by the following hypothesis made by Michael Stubbs on the possibilities for looking 'upwards' from the lexicogrammar to identify patterns of meaning in the discourse semantics:

> Vocabulary and grammar provide us with the potential and resources to say different things. But often this potential is used in regular ways, in large numbers of texts, whose patterns therefore embody particular social values and views of the world. Such discourse patterns tell us which meanings are repeatedly expressed in a discourse community. [1]

Method: The two discourse communities whose language I wished to study and compare were (i) self-styled green business corporations and (ii) environmentally-focussed NGOs which were sceptical to the sustainability claims of such companies. In the paper I describe the requirements for 'membership' in the two discourse communities. Then I outline the design and construction of the two corpora which provide their respective lexicogrammars. I describe the procedures developed for generating what I describe as 'contextualised concordance' reports on a small selection of keywords common to both corpora. These take advantage of the functionality available in Wordsmith Tools – the only electronic corpus tool that I utilised in the project.

Results: I present some of my 'contextualised concordance' reports and explain my interpretive reading of the usage of these keywords in the two corpora. I argue that these reports demonstrate a consistent, markedly different usage of keywords in the two corpora.

Conclusions: Further, I suggest that an 'aggregation' of these differences across a selection of semantically-related keywords points to considerable divergences in the ways in which these two discourse communities think about the relationship between business and the natural environment. However, the results and their conclusions need to be tempered by several assumptions and simplifications which had to be made in the methodology. These are presented for criticism and as a possible agenda for future empirical work in this area.

Returning to the motivating hypothesis for the project, the patterns of wording and the patterns of meaning which I identify are entirely consistent with the expected 'world views' of the two communities. This, I suggest, provides empirical support for Stubbs' claim.

Reference

[1] Michael Stubbs, Text and Corpus Analysis: Computer-assisted Studies of Language and Culture, (Oxford: Blackwell, 1996), 158

Hesitation markers occurring in connection with adjective modification in Swedish learners' and native English speakers' spoken English

Viktoria Börjesson
University of Gothenburg

This study is part of a thesis project on the use of reinforcing and attenuating modifiers of adjectives in Swedish NNS and NS English. It describes and discusses hesitation occurring in connection with adjective modification, based on occurrences in the Swedish component of the *Louvain International Database of Spoken English* (LINDSEI), and the native speaker *Louvain Corpus of Native English Conversation* (LOCNEC).

The hesitation markers discussed are repetitions (*it's very it's very different*), silent and filled pauses, discourse markers (e.g. *sort of, you know, like*) and false starts, most often followed by rephrasing, such as *it's very .. they need eh teachers now.* A comparison is made based on whether the repetition, pause, etc. is produced before or after the modifier, and whether the modifier and adjective are separated by hesitation or are produced as one unit, the hypothesis put forward being that native speakers are more likely than learners to produce chunks of words whereas learners would treat modifier and adjective as separate units, relying on the previous finding by other scholars that prefabricated patterns are more common in native speaker speech.

The result shows for example that, in comparison with native speakers, Swedish learners hesitate more often before the adjective, using pauses (both silent and filled) and discourse markers. Swedish learners also rephrase more frequently, and leave the modifier 'dangling', without any adjective following. It may thus be assumed that modifiers and adjectives are less commonly stored as chunks in the memory of learners. However, the difference observed can also be referred to the fact that the majority of the modifiers are highly frequent and versatile (*very, really, quite, so, a bit*, etc.), thus easily retrievable from memory, whereas finding the accurate adjective may turn out to be a more complicated task.

Variability in the use of prepositions as an indicator of direct transitivization in Present-Day American English

Marcus Callies
Johannes Gutenberg University Mainz

The use of prepositions in various grammatical patterns has undergone significant changes in the (more recent) history of English. For instance, in several verbs prepositions like *against* or *from* are increasingly omitted in favour of direct transitivization, as exemplified below:

(1) On March 5, 3000 people marched through Hackney to <u>protest Ø the raid</u>. (BNC HSL)
(2) Without another word, he <u>departed Ø the room</u> and returned to the study where Luther was waiting for him. (BNC FPK)

Two semantic classes of verbs have been investigated in some detail (Hundt, 1998, 1999; Rohdenburg 2009): antagonistic verbs (*appeal, battle, fight, protest*) and verbs of leaving (*depart, escape, flee, resign*). The decrease of prepositional objects after these verbs has been interpreted in line with an ongoing tendency in the history of

English to functionally expand the category of the direct object at the expense of prepositional phrases in particular, with American English being assumed to be further advanced in this development since the preposition-less variant has become the preferred option in the course of the 20th century (Rohdenburg 2009: 200). Other verbs that are said to exhibit the same phenomenon are mentioned only sporadically in the literature (Algeo 1988, 2006) but have not yet been examined systematically on a broader empirical basis.

This paper presents evidence from corpora of American English suggesting that several other verbs are also being affected by this development. It appears that with these verbs, prepositions are semantically redundant, not adding significantly to the meaning of the verb, and at the same time grammatically and functionally omissible. The paper argues that this development has to be seen in line with a larger set of what may be described as erosion processes due to verbal economy rather than informality in Present-Day English. For instance, prepositions are also frequently omitted in complex noun-noun sequences that lead to a greater lexical density of press texts in particular, typical of American English usage. American English has been claimed to be the more dynamic variety in leading the development in such erosion processes, favouring formally less explicit or simpler variants over more complex and explicit ones (e.g. Rohdenburg 2009). The current study thus provides further empirical support for the recently advanced hypothesis that American English grammar "shows a more marked tendency to dispense with function words that are semantically redundant and grammatically omissible", and that this "trend towards grammatical economy ties together an array of otherwise unrelated phenomena in the complementation system" (Rohdenburg & Schlüter 2009).

References

Algeo, John (1988), "British and American grammatical differences", *International Journal of Lexicography*, 1(1), 1-31.

Algeo, John (2006), *British or American English? A Handbook of Word and Grammar Patterns*. Cambridge: CUP.

Hundt, Marianne (1998), *New Zealand English Grammar – Fact or Fiction? A Corpus-Based Study in Morpho-Syntactic Variation*. Amsterdam: Benjamins.

Hundt, Marianne (1999), "The press sections of standard one-million-word corpora", in Diller, Hans-Jürgen, Erwin Otto and Gert Stratmann (eds.), *English Via Various Media* (anglistik&englischunterricht 62). Heidelberg: Winter, 155-177.

Rohdenburg, Günter (2009), "Nominal complements", in Rohdenburg, Günter & Julia Schlüter (eds.), *One Language, Two Grammars? Differences between British and American English*. Cambridge: CUP, 194-211.

Rohdenburg, Günter & Julia Schlüter (2009), "Introduction", in Rohdenburg, Günter & Julia Schlüter (eds.), *One Language, Two Grammars? Differences between British and American English*. Cambridge: CUP, 1-12.

Introducing the Corpus of Academic Learner English (CALE)

Marcus Callies, Ekaterina Zaytseva
Johannes Gutenberg University Mainz

This work-in-progress report introduces the *Corpus of Academic Learner English* (CALE), a learner corpus for the detailed quantitative and qualitative description of advanced learner varieties as to written academic English. In recent years, SLA research has seen an increasing interest in advanced stages of acquisition and questions of near-native competence, and corpus-based research into learner language has contributed to a much clearer picture of advanced interlanguages. There is evidence that learners of various native language backgrounds have similar problems and face similar challenges on their way to near-native proficiency. For example, advanced learners still struggle with the acquisition of optional and/or highly L2-specific linguistic phenomena, often located at the interfaces of linguistic subsystems. Also, in academic writing, many of their difficulties appear to stem from a lack of understanding of register-specific rules, or from a lack of practice, rather than as a result of interference from L1 academic conventions. Due to these similarities, we refer to the interlanguage of these learners as advanced learner varieties (ALVs). Despite the growing interest in the concept of advancedness the field is still struggling with 1) a definition and clarification of the concepts "advanced learner" and "nativelikeness", 2) an in-depth description of ALVs, especially when it comes to learners' acquisition of optional and highly L2-specific phenomena in all linguistic subsystems, and 3) the operationalization of such a description in terms of criteria for the assessment of advancedness.

While existing learner corpora, such as the *International Corpus of Learner English*, include learner writing of a general argumentative, creative or literary nature, CALE is designed to comprise a range of academic genres produced by EFL-learners in a university setting across several disciplines. Thus, CALE may be conceived of as what has recently been termed a Language for Specific Purposes learner corpus, containing discipline- and genre-specific texts (Granger & Paquot, forthc.). Possible native-speaker control corpora for CALE are the *Michigan Corpus of Upper-Level Student Papers* (MICUSP, Römer & Brook O'Donnell, in prep.) or the *British Academic Written English* corpus (BAWE, Alsop & Nesi, 2009).

This report outlines the corpus design (classification of text types, annotation system) and the specific research program that CALE will be used for, i.e. the investigation of various factors that determine patterns of lexico-grammatical variation in ALVs, ensuring that the findings can be applied to improve teaching in academic writing classes at the advanced level.

References

Alsop, S. & Nesi, H. (2009), "Issues in the development of the British Academic Written English (BAWE) corpus", *Corpora*, 4:1, 71-83.

Granger, S. & Paquot, M. (forthcoming), "Language for Specific Purposes Learner Corpora", in Upton, T.A. & Connor, U. (eds*.), Language for Specific Purposes. The Encyclopedia of Applied Linguistics*. New York: Blackwell.

Römer, U. & Brook O'Donnell, M. (in preparation). From student hard drive to web corpus: The design, compilation, annotation and online distribution of MICUSP.

Alternatives for causal questions: *why*, *what … for*, and *how come*

Claudia Claridge
University of Duisburg-Essen

How come and *what … for* are lexicalized alternatives for *why* in English, but neither has received much attention in the literature. The present contribution will investigate their usage in American English. The data will be taken from the *Corpus of Contemporary American English* (COCA) and the *Corpus of Historical American English* (COHA).

While forms involving *what* and *for* have a long history in English, *how come* is a more recent form and seems to be of American origin – which is why this variety is chosen here. While *why* and idiomatic *how come* are easy to search for, *what … for* needs considerable disambiguation to sort out non-relevant items like *what you're getting yourself in for?, what are you going to be for?* – whose presence might inhibit a greater use of the causal variant.

Why is by far the most frequent form (980 instances per million), with *how come* (15.8 per million) and *what ... for* (2.4 per million) representing clear minority options. All of them are more frequent in speech(-related) contexts, but *how come* and *what … for* are most frequent in fiction. In this respect, it is interesting to investigate whether these forms serve as special orality/informality markers in literature.

Other aspects to be looked at include:

- Are the forms distinguished regarding their semantic specialisations, namely purpose (*what ... for*), cause (*how come*) or both (*why*), as postulated by Zwicky and Zwicky (1973), but denied by Huddleston and Pullum (2002)?
- What is the syntactic behaviour of these forms like, e.g. how are their realisations distributed across main and subordinate clauses, what is their scope in complex clauses (e.g. *what ... for*: whole sentence, *how come*: main clause only)?
- Is there any semantic, syntactic or stylistic development noticeable? *How come* only becomes frequent from the 1940s onwards, which might go hand in hand with a functional expansion. *What ... for*, in contrast, shows stable frequencies according to COHA, which may go along with unchanging usage.
- *What ... for* presents a special case insofar as it has variants, namely bare *what for?* and *for what (...)?*. All of these are attested historically, but it has been claimed that *for what* is not possible in modern English (Quirk et al. 1985). COCA, however, does yield potential instances, such as: *The county sheriff would have to arrest me. – For what? Not indecent exposure.* Thus the frequency and contexts of these variants will also be investigated.

References

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

Bolinger, Dwight. 1970. "The lexical value of it." *Working Papers in Linguistics, University of Hawaii*, 2 (8), 57-76.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language.* Cambridge: CUP.

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

Zwicky, Arnold M. & Ann D. Zwicky. 1973. "*How come* and *what for*." In: B. Kachru et al. (eds.), *Issues in Linguistics. Papers in Honor of Henry and Renée Kahane.* Urbana: University of Illinois Press, 923-933.

Contiguous adjectives of size: 'great big', 'tiny little', and less frequent pairings and triplets

Stephen James Coffey
Università di Pisa

In the proposed presentation, I report on a corpus study of the phenomenon whereby two, and sometimes three, adjectives of size are used together in prenominal position. The adjectives in question denote the basic notions of 'bigness' and 'smallness', and this use of more than one adjective can most easily be viewed as a way of emphasizing the quality being denoted. Examples of phrasal pairs are the fairly frequent ones mentioned in the title (*great big*, *tiny little*), and less frequent ones such as *big gigantic*, *massive big* and *tiny wee*. An example of a triplet is *huge great big*. Short corpus (BNC) contexts for these particular phrases are: 'It's a big gigantic screw'; 'They've all got massive big places'; 'theres's a tiny wee corner at the top'; and 'it's a huge great big muscle'.

I am not familiar with previous studies which describe this phenomenon. Notably, there is no reference to it in Bolinger's very thorough account of 'intensification' (1972). As a precise lexico-semantic-grammatical pattern, it is not particularly frequent, and therefore does not find its way into descriptive grammars. In terms of the individual adjective combinations, just a few are used commonly enough to show up in frequency-based corpus analysis and lexicographical description (including the occasional lexicalized reduplicative, e.g. *teeny weeny*).

In order to document modern usage, a list of basic 'size' words was drawn up, and the various resulting word combinations were then looked for individually in a number of corpora. The conference presentation will report on: which 'types' were found (to date, more than 50 combinations); corpus frequencies; text types; any geographical dependency (Br.Eng. vs N. Am. Eng.); any tendencies of specific adjectives to be in first or second position; and the nature of the nouns that are modified.

Specific comments will also be made on cases where there is an intervening comma (e.g. 'these big, massive sounds'), some examples of which are in predicative position.

As a supplement to the description, and to delimit the phenomenon, some data will also be provided regarding: 1) adjective combinations in which one of the adjectives contains denotative or connotative semantic features in addition to the central notion of 'size' (e.g. *big hefty*, *dinky little*); 2) analogous repetitive combinations from other semantic fields (e.g. *filthy dirty*).

Reference

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

The present perfect in World Englishes: A corpus-based study

Peter Collins, Xinyue Yao
University of New South Wales

A number of recent studies have investigated the uses of the present perfect in English, its diachronic development, and the variation in usage between British and American English. More recently, researchers have focused on non-standard uses of the present perfect in simple preterite contexts in Australian and New Zealand English, interpreting them as possible signs of diachronic change. No comprehensive study has been carried out to investigate regional variation in the distribution of the present perfect beyond Englishes of the 'Inner Circle'.

This paper reports the findings of a study that compares and contrasts the distribution of the present perfect in ten regional varieties, encompassing both Inner and Outer Circle Englishes. The International Corpus of English (a set of parallel one-million-word corpora) was selected as the primary source of data. A macroscopic quantitative approach was adopted to examine the overall frequencies of the present perfect and its chief competitor the simple preterite in the selected varieties. Attention was also given to differences in their patterning across a range of spoken and written genres. A further qualitative analysis was carried out to examine the discourse and pragmatic functions of the present perfect and its interaction with various temporal expressions. In analyzing differences across the various Englishes, consideration was given to the historical status of the present perfect and the relationship between its distribution and such ongoing processes as 'Americanization' and 'colloquialization'.

Reappropriation and its effects: A corpus-based analysis

Anne Curzan
University of Michigan

Linguistic reappropriation, as a process of semantic change, has received relatively little scholarly attention. It is often addressed with a sentence or two in a broader discussion of semantic change, sometimes linked to amelioration, with a note that in this case a group is consciously "taking back" a word. Galinsky et al. (2003: 222), one of the two most substantive studies of the process in the past decade (see also Brontsema 2004), define r*eappropriation* as "the phenomenon whereby a stigmatized group revalues an externally imposed negative label by self-consciously referring to itself in terms of that label." Previous studies have, importantly, teased apart the goals of reappropriation (e.g., value-reversal vs. neutralization) and different stances about the nature of lexical meaning. There have, however, been no previous systematic attempts to track the effects of reappropriation efforts on English usage. This paper works from two central case studies: the feminist reappropriation of the word *woman*, to replace *lady*, beginning in the nineteenth century; and the gay and queer community's reappropriation of the words *gay* and *queer*, beginning in the second half of the twentieth century. Data from the Corpus of Historical American English (COHA) reveal not only critical shifts in frequency but also revealing trends in collocational patterns. The study also analyzes patterns in register distribution, working from a comparison of data from the Corpus of Contemporary American English (COCA) with data from COHA, in order to describe the diffusion of more conscious language change of this type. Historical studies of English have often downplayed conscious efforts at language change/reform, focusing instead on what is typically described as "natural" language change—a dichotomy productively challenged by Deborah Cameron's concept of verbal hygiene. This study works from the fundamental premise that to fully describe the ongoing history of English, we must better understand the mechanisms and effects of socially motivated, conscious (and sometimes remarkably successful) attempts to alter usage, in this case specifically of identity terms.

References

Brontsema, Robin. 2004. A Queer Revolution: Reconceptualizing the Debate Over Linguistic Reclamation. *Colorado Research in Linguistics* 17.1: 1-17.

Galinsky, Adam D., Kurt Hugenberg, Carla Groom, and Galen Bodenhausen. 2003. The Reappropriation of Stigmatizing Labels: Implications for Social Identity. *Research on Managing Groups and Team*s 5: 221-256.

Mom and Dad but *Men and Women*: The sequencing of sex-dependent noun pairs

Doris R. Dant
Brigham Young University

One aspect of gender bias in language that has received comparatively little attention is how pairs of sex-dependent nouns (e.g., *men and women*) are sequenced in actual practice. The American Psychological Association style manual recommends either altering the traditional order of such pairs or alternating the position of the nouns. Such a practice should result in considerable variability in positioning. *Biased* and other guides to nonsexist language echo this recommendation to avoid privileging one noun over the other by always placing it first; some evidence shows that the first position is better accessed. To determine how 76 pairs of sex-determined nouns have been sequenced in the United States, I queried the Corpus of Contemporary American English (COCA) and the Corpus of Historical American English (COHA). The results indicate that US English has been quite resistant to this recommendation and that the feminism movement does not account for most of the pairs where the female noun regularly occurs first. The sequencing of only three pairs appears to be in free variation and that in just some of the registers. In ten of the pairs I investigated—nine of which show family relatedness (e.g., *[aunt] and [uncle]*)—the female noun is in first position 60 percent or more of the time. For six of those relationship pairs, this positioning gained dominance in at least two registers even before second-wave feminism began calling attention to such matters. *Ladies and gentlemen*, the seventh pair, has been idiomatic since the 1400s and has been the preferred sequence in the United States at least since 1810 (the earliest date in COHA). Twenty pairs of nouns range from the male noun always being in first position to the female noun appearing first only 39 percent of the time in at least one register. The largest category consists of parallel terms that for various reasons are rarely or never paired, such terms as *[poet] and poetess/poetesses, [widower] and [widow]*, and [*confidant] and [confidante]*. One reason is that the feminine form is rarely used in contemporary American English. Another is that contexts in which the parallel terms might be paired are relatively sparse. In all categories—shorter noun first, nouns of equal length, longer noun first—pairs with the male noun in first position outnumbered those with the female noun first. This finding indicates that the privileging of the male noun frequently overrides any preference for placing the noun with fewer syllables in first position.

Approximating devices in English and French business news reporting: more or less the same?

Sylvie De Cock, Diane Goossens
Université catholique de Louvain

A corpus-driven study of number approximations (e.g. *about 450*) in English business news reporting brought to light the wide grammatical and semantic diversity of devices that can be used to approximate numbers expressing quantities in this genre (Goossens and De Cock 2010). The investigation also uncovered some of the preferred collocational patterns in which combinations of approximators and numbers tend to occur (e.g. *Rio's stockmarket value* <u>*hovers*</u> *around $35 billion*). The aim of this paper is to analyze combinations of numbers and approximators in a corpus of French business news reporting and to explore the similarities and differences between the devices used to approximate numbers in English and French. The data used in the study include two comparable 500,000-word corpora of business news reporting: the Business English News corpus and the recently compiled Business French News (Centre for English Corpus Linguistics). The analysis of number approximations in French is conducted using a similar corpus-driven method as that adopted in Goossens and De Cock (2010), i.e. taking concordances of numbers automatically retrieved from a part-of-speech tagged version of the corpus as a starting point to explore the various items (e.g. words, parts of words or punctuation) used around these numbers to approximate quantities. The contrastive study sets out to compare not only the extent to which numbers expressing quantities are approximated in French and English but also the grammatical and semantic variety of approximating devices that can be found in the two corpora. In other words, the paper seeks to answer the following questions: (1)'Does English tend to use more imprecision when expressing quantities than French?', (2) 'Do French and English display (dis)similar preferred grammatical categories when approximating numbers in business news reporting (e.g. adverbs, prepositions, verbs, prepositional phrases, prefixes or suffixes)?', (3) 'Do the two languages exhibit (dis)similar semantic tendencies when expressing approximation of quantities in business news reporting, i.e. do they favour approximators expressing a minimum amount (e.g. *au moins 30%, larger than 30 inches*), a maximum amount (e.g. *fewer than 300, tout au plus 20*), an interval (e.g. *from one to two, entre 2 et 5*) or an amount which is equal to more or less the number used (e.g. *plus ou moins 100£, around 25%*)?', and (4) How do combinations of approximators and numbers in the two corpora compare in terms of the preferred company they keep?.

Pragmatic profiling of business corpora: speech act tagging

Rachele De Felice, Svenja Adolphs
University of Nottingham

This presentation describes the first stage of a two-year project which applies corpus analysis and natural language processing techniques to create a comprehensive profile of the pragmatic characteristics of spoken, written, and email Business English. In particular, we discuss the feasibility of extending a speech act tagger developed for workplace emails written by nonnative speakers (cf. De Felice and Deane 2009) to corpora of native-speaker Business English, such as the Wolverhampton Business English corpus (10 million words of written text), the Enron email corpus (Berry, Browne, & Signer, 2007), and the Cambridge and Nottingham Spoken Business English Corpus (1 million words). The speech act tagger has been designed to recognise speech acts typical of email communication such as requests, orders, and commitments. The challenges encountered in applying and adapting the tagger to the spoken and written data highlight how these forms of communication differ from email language, helping us draw a picture of pragmatic variation across the three types, for example in the differing frequencies of particular speech acts, or in the way they are introduced and formulated. Sentence-level speech act tagging is the first step towards a more detailed analysis of the different types of speech acts, which will consider their lexical and grammatical characteristics, such as which verb forms are most common, or which lexical items feature most often as subjects. Understanding how speech acts are typically formulated, and how they contribute to the discourse structure of business communication, are key elements for the description of the different forms of Business English. This information is of particular benefit to those unfamiliar with the conventions of this type of language, be they non-native speakers or those just entering the workforce for the first time.

References
Berry, M., Browne, M., & Signer, B. (2007). 2001 Topic Annotated Enron Email Data Set. Philadelphia: Linguistic Data Consortium.
Rachele De Felice and Paul Deane (2009). Identifying speech acts in emails: Business English and nonnative speakers. Corpus Linguistics Conference, Liverpool, UK

The return of the prefix? New verb-particle combinations in blogs

Stefan Diemer
Technical University Berlin

This paper will explore how verb-particle combinations, one of the most productive segments of English word-formation, have changed with the advent of online real-time short communication forms such as blogs or their more sophisticated social networking or microblogging varieties like Facebook and Twitter.

One of the main trends in the development of English is the long and seemingly unstoppable rise of verb-adverb combinations and the accompanying decline of the prefixes, especially during the Middle English period. As a result, modern English has only very few productive verb prefixes left, in contrast to other languages such as German, where prefixed verbs are much more common and remain productive. This comparatively stable situation may be changing.

Following up on earlier research (Diemer 2008 and 2010), evidence will be presented that the long-term decline of prefixed verb forms has been stopped and even partly reversed by these new forms of communication, which seem to facilitate the use of previously non-standard prefixed verbs like *inbe*, *oncome* and *atstand* in both native and non-native English blogs. Selected examples will be discussed on the basis of an extensive corpus of blogs and contrasted with existing forms in German.

It will be argued that the main reasons for this change are facilitation of syntax, need for innovation in specialized and peer group communication, analogy formation and the influence of other languages on English.

Select bibliography
Abraham, Lee B. & Lawrence Williams (eds.). 2009. Electronic discourse in language learning and language teaching. Amsterdam: John Benjamins.
Bansal, Nilesh & Nick Koudas. 2007. Searching the Blogosphere. In Proceedings of the 10th international Workshop on Web and Databases. Beijing: WebDB 2007.
Crystal, David. 2008. Txtng: the Gr8 Db8. Oxford: Oxford University Press.
Diemer, Stefan. 2008. Die Entwicklung des englischen Verbverbandes – eine korpusbasierte Untersuchung. Habilitationsschrift (professorial thesis). Berlin: TU Berlin.
Diemer, Stefan. 2010. „It’s all a bit upmessing. Non-standard verb-particle combinations in blogs.” In: Saarland Working Papers in Linguistics 3 (2009): 35-56.
Giltrow, Janet & Dieter Stein (eds.). 2009. Genres in the internet. Amsterdam: John Benjamins.

Hiltunen, Risto. 1983. The decline of the prefixes and the beginnings of the English phrasal verb. Turku: Turun Yliopisto.

Unreal conversation taking place in unreal time: Spoken style in fictional scripted television language

Stefanie Dose
Justus Liebig University Giessen

Biber et al. (1999: 1041-1051) have identified a number of 'discourse circumstances' which influence the linguistic choices speakers make in conversation, e.g. that "conversation takes place in shared context" (1042), "conversation is interactive" (1045), and "conversation takes place in real time" (1048). A multitude of grammatical features reflect these circumstances and are thus characteristic of a 'spoken grammar,' e.g. performance phenomena, discourse markers, and other features which might simply display quantitative differences compared to written registers.

This study focuses on one of these discourse circumstances, viz. "conversation takes place in real time", which means that speakers constantly face the pressure of planning and producing utterances "on the fly" (Biber et al. 1999: 1048). This is typical of natural, spontaneous conversation – but what about the case of fictional scripted television language? On the one hand, conversation obviously does not take place in real time here because the dialogues are constructed beforehand by the scriptwriters; thus they do not have to be planned in real time. On the other hand, the utterances are still produced in real time by the performing actors. On the basis of CATS (a Corpus of American Television Series), consisting of 160,592 words of spoken language from four contemporary television series, this study investigates how exactly these altered discourse circumstances are manifested in terms of grammar and where this complex variety can be situated on the continuum between speech and writing.

For instance, a feature which is closely associated with the spoken language is the use of contracted verb forms (e.g. Biber 1988). The present study thus looks at 'personal pronoun + verb contraction' structures as indicators of spoken style. While previous analyses of other spoken features connected to the real time context (e.g. hesitators, discourse markers) revealed quite some discrepancies between fictional scripted speech and spontaneous speech (cf. Dose forthcoming), it will be shown that the use of verb contractions in television language is strikingly similar to naturally occurring speech. A quantitative analysis with CATS indicated e.g. that contractions with 's/'re/'m are in fact even more frequent than in a corpus of naturally occurring speech, i.e. c. 25,000 compared to c. 20,000 instances pmw (cf. Biber et al. 1999: 1062).

Scripted television language is a complex register conditioned by quite 'mixed-up' discourse circumstances. Investigation of its oral/literary status therefore always has to systematically consider a variety of factors. This study seeks to contribute to the exploration of a field which, with a few exceptions (e.g. Quaglio 2009), still lacks comprehensive theory and description.

References

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

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

Dose, Stefanie. forthcoming. “Flipping the script: A Corpus of American Television Series (CATS) for corpus-based language learning and teaching”. In Magnus Huber and Joybrato Mukherjee (eds.): *Corpus Linguistics and Variation in English: Focus on Non-native Englishes. Proceedings of ICAME 31*. [eVARIENG: Studies in Variation, Contacts and Change in English]

Quaglio, Paulo (2009): *Television Dialogue: The Sitcom* Friends *vs. Natural Conversation*. Amsterdam: John Benjamins.

Using translation corpora to explore synonymy and polysemy

Thomas Egan
Hedmark University College

Polysemy and synonymy may be viewed as two sides of the same coin, in that in both cases we are faced with one-to-many form-meaning relationships. The purpose of the paper is to show how translation corpora can be mined to shed light on such one-to-many relationships. With respect to synonymy, the greater the degree of semantic overlap there is between two lexemes or constructions in language A, the more difficult it should be to predict the original forms given their translations into language B. As for polysemy, it is hypothesised that putatively different senses of a lexeme or construction are more likely to be translated differently than similar senses. The more similar the translation equivalents in language B of two senses, the closer the relationship is likely to be between those two senses in the polysemous network of the lexeme or construction in language A (see Egan forthcoming).

The data for this paper comprise all tokens in the English Norwegian Parallel Corpus (see Johansson 2007) of the two near-synonymous verbs *begin* and *start* and of the multi-polysemous preposition *at*, of which there are more than 2,500 tokens in the ENPC and for which there is no one single Norwegian equivalent. I investigate how the translations of these forms into Norwegian can aid us in ascertaining the extent to which the former pair may be said to be synonymous and in tracing the polysemous semantic network of the preposition. I show that *begin* and *start* are to all intents and purposes synonymous in some, but not all, syntactic frames. I also show that, with one significant exception (the Perception sense, instantiated by *look at*) the various senses of *at* cluster into two main semantic sub-networks.

References

Egan, T. (forthcoming) *Through* seen through the looking glass of translation equivalence: a proposed method for determining closeness of word senses. In *Corpus linguistics: Looking back – moving forward*. S. Hoffman, P. Rayson and G. Leech (eds.). Amsterdam: Rodopi.

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

Language and culture: on evidence from language corpora about the development of cultural differences between English-speaking countries

Johan Elsness
University of Oslo

Having studied lexical frequencies in the Brown and LOB corpora, Leech and Fallon (1992) venture the following conclusion:

> Wrapping up the whole analysis ... in one wild generalization, we may propose a picture of United States culture in 1961 – masculine to the point of machismo, militaristic, dynamic and actuated by high ideals, driven by technology, activity and enterprise – contrasting with one of British culture as more given to temporizing and talking, to benefitting from wealth rather than creating it, and to family and emotional life, less actuated by matters of substance than by considerations of outward status. (Pp. 44-45)

This conclusion was based on frequency differences from a large variety of cultural and other domains: arts and education, sports and travel, administration and politics, law and military, religion and personal reference. The latter especially involved gender distinctions, where the predominance of male terms was found to be even more overwhelming in Brown.

In this paper results are presented from a wider comparison, taking in the two Freiburg updates of Brown and LOB from 1991/1992, and also the British National Corpus (BNC) (mostly late 1980s and early 1990s) and the Corpus of Contemporary American English (COCA) (1990-2010). The general trend is shown to be that terms which were overrepresented in Brown are still more frequent in the American material from a generation later but less markedly so; and a corresponding trend is noticeable for terms which were overrepresented in LOB. In the case especially of religious references a notable reversal of the downward trend in AmE is recorded in the material covering the last couple of decades.

Comparison is made throughout between different text categories, in the case of the BNC and COCA including the fundamental opposition between speech and writing. In the BNC male predominance is recorded in both men's and women's speech and in men's writing, only female writers displaying a (slightly) higher frequency of female references.

In some cases longer-term trends are reported from the Time Magazine Corpus (1923-2006), and also results from the Australian Corpus of English (1986) and the Wellington Corpus of Written New Zealand English (1986-1990).

Overall results are seen to reflect some pretty fundamental cultural shifts within the English-speaking world. In many cases there is evidence of a general cultural convergence as we all became citizens of the global village.

Reference

Geoffrey Leech & Roger Fallon (1992), 'Computer Corpora – What do they tell us about culture?', *ICAME Journal* 16: 29-50 [Reprinted in Geoffrey Sampson and Diana McCarthy (eds.) (2004), *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum. Pp. 160-173.]

Interrelations between visuals and writing in news reporting: a multimodal corpus-based study

Roberta Facchinetti
University of Verona

News media studies have so far rarely dealt with texts and their related images from a multimodal point of view, thus failing to highlight the semantic and pragmatic significance of their interplay (Jewitt and Kress 2003, Kress and van Leeuwen 2006).

The present research aims at bringing to the foreground such interrelations in newspaper discourse (Popp and Mendelson 2010), by means of a case study; specifically, the focus will be on the coverage by *The New York Times* of the birth and development of the European Union. To do so, I have compiled a corpus of the front pages from *The New York Times*, published between 1945 and 2009. The data have been screened manually in order to select all news reports dealing with the process of European integration.

The analysis of the corpus focuses particularly on the interplay between the 'syntactic implicitness' (Messaris and Abraham 2001: 220) of the cover pictures, the related lexical context (captions and headlines), the associated body texts and also their position in the front page layout. The results testify to the fact that visuals and writing are distinct but equally significant and interrelated elements of the news-making process; furthermore, their interrelation strongly adds to the factual and evaluative information of the journalistic pieces and also, in certain cases, helps framing and articulating ideological messages. Specifically, the multimodal analysis has highlighted the following:

1) no regular direct alignment between visuals and writing: visuals do not always mirror what is reported in writing (particularly in headlines); this is frequently the case when the piece highlights problematic aspects of the EU making process;
2) gradual, steady change in presenting the news on European integration by the *New York Times* throughout the decades under scrutiny, particularly with reference to (a) the positioning of the pieces within the front page (b) the role played by the United States, and (c) the visuals selected.

Bearing in mind the pivotal role of new media in present-day journalism, the paper also advocates the importance of exploiting multimodal corpora in certain contexts of study; indeed, the present analysis testifies to the importance of interrelating visuals and writing in the analysis of journalistic discourse, since only by focusing on such interplay is it possible to identify special features of the news-making and news-delivery process.

References

Kress G.R. and T. van Leeuwen (2006) *Reading Images: The Grammar of Visual Design*, 2nd ed., London Routledge.

Jewitt C. and G.R. Kress (eds) (2003) *Multimodal Literacy*, New York: Peter Lang.

Messaris P. and L. Abraham (2001) "The role of images in framing news stories", in S. Reese, O.H. Gandy Jr., A.E. Grant (eds.) *Framing Public Life. Perspectives on Media and our Understanding of the Social World*. Mahwah, N.J.: Lawrence Erlbaum, pp. 215-226.

Popp R.K. and A.L. Mendelson (2010) "'X'-ing out enemies: *Time* magazine, visual discourse, and the war in Iraq", *Journalism* 11(2): 203-221.

Exploring lexical gravity within a multi-faceted approach to the study of collocation: preliminary proposals

Adriano Ferraresi
University of Naples “Federico II” / University of Bologna

According to Wray (2002:66), formulaic language is “not a single and unified phenomenon”, and “several baselines” are needed to account for what is (or is not) formulaic. The same could be said of the more specific notion of lexical collocation: depending on the definition of collocation, different criteria are invoked in the literature to assign “collocational status” to sequences of two (or more) words, e.g. frequency of co-occurrence, idiosyncrasy, saliency in the mind of native speakers, etc. The present paper describes the first steps of a project which aims at taking into account such complexity in evaluating statistical measures for collocation discovery.

This work-in-progress report describes a small-scale experiment concentrating on a specific collocational measure, namely *lexical gravity* (Daudaravičius and Marcinkevičienė 2004). Unlike other measures, which are based on words’ joint frequencies and their individual token counts only, lexical gravity also takes into account the frequency of co-occurrence of types, i.e. given two words X and Y, it considers how many different *types* co-occur with X in the position of Y (and vice versa). The measure has been used in Gries (2010), and Gries and Mukherjee (2010), but has not been analysed per se as a measure of collocational strength and compared to better-established ones. Our aim at this stage is to present the first phase of its evaluation, focusing on what Nessalhauf (2004) calls the “statistical dimension” of collocations. Lexical gravity values are calculated for adjective-noun sequences extracted from a specialised corpus of English consisting of webpages of British and Irish universities (Bernardini et al. 2009) and are compared to bare frequency of co-occurrence and the widely used Mutual Information and log-likelihood by means of rank correlation tests. Manual inspection of the lists is carried out to check to what the extent the measures can discriminate between salient and less salient pairs.

In the conclusions, we discuss further evaluation steps i.e. a) using dictionary evidence to check which measure corresponds better with “lexicographic salience”, and b) tapping into what Partington (1998) calls the “psychological” dimension of collocations through e.g. questionnaires and collocativity judgement tests. The advantages and potential shortcomings of each method will be briefly pondered. We thus hope to stimulate discussion on the nature of the evidence needed to provide a clearer picture of the properties of association measures whilst mirroring the complex nature of lexical collocation.

References

Bernardini, S., A. Ferraresi and F. Gaspari. 2009. “Institutional English in Italian University websites: the acWaC corpus”. Paper presented at *Corpus Linguistics 2009*, University of Liverpool.

Daudaravičius, V. and R. Marcinkevičienė. 2004. “Gravity counts for the boundaries of collocations”. *International Journal of Corpus Linguistics* 9(2). 321-348.

Gries, S. Th. 2010. “Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora”. In *Proceedings of Corpus Linguistics 2009*, University of Liverpool.

Gries, S. Th. and J. Mukherjee. 2010. “Lexical gravity across varieties of English: an ICE-based study of n-grams in Asian Englishes”. *International Journal of Corpus Linguistics* 15(4). 520-548.

Nesselhauf, N. 2004. *Collocations in a learner corpus*. Amsterdam: Benjamins.
Partington, A. 1998. *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam: Benjamins.
Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Exploring the dialogism of academic discourse: appraisal in a multimodal corpus of medical research articles

Daniel Lees Fryer
University of Gothenburg

Academic discourse is dialogic, and the way in which a researcher engages with other voices in the discourse is an integral part of the social practice of communicating research. In this work-in-progress paper, I will discuss how this dialogism is realized in medical research discourse, by applying the systemic-functional framework of appraisal to a multimodal corpus of medical research articles. Specifically, I will present how the corpus has been compiled and annotated according to the system networks of engagement and graduation. I will also present preliminary findings of the discourse-semantic features identified, the probabilities of their being selected, and their distributions across the texts. Some of the challenges involved in annotating nonverbal and multimodal elements, e.g., figures and tables, as well as general challenges related to corpus-based application of the appraisal framework will also be discussed.

Assessing social responsibility: a corpus-based analysis of Appraisal in BP and Ikea's social reports

Matteo Fuoli
University of Trento

Saturated markets, heightened competition and the emergence of new forms of critical consumption compel multinational corporations to invest increasing resources in the implementation and promotion of principles of ethical business. In 'Sustainability Reports', companies account for and assess their performance across the 'triple bottom line' (environment, society, profit).

Despite the wealth of research on evaluation (see Hunston and Thompson 2000), few studies have concerned the genres of business communication (see e.g. Malavasi 2007, 2008). The present work aims at partially filling this gap by applying the Appraisal theory (Martin 1995, 2000; Martin and White 2005; Macken-Horarik and Martin 2003; White 2001) to the analysis of evaluation in a small corpus, comprised of BP and Ikea's 2009 sustainability reports (total word count: aprox. 55000 tokens). Based on the assumption that evaluation plays a fundamental role in the rhetorical 'texturing' of social and institutional identities (Fairclough 2003), the analysis aims to show how these two companies use evaluative resources to represent themselves and to construe their relationship with their stakeholders.

The analysis is quantitative and focuses on the Appraisal systems of Attitude and Engagement, the former concerning the linguistic expression of affect and attitudes, the latter encompassing a wide range of resources that have been studied under the headings of 'evidentiality' (Chafe and Nichols 1986), 'hedging' (Hyland 1996), 'modality' (Hoye 1997, Palmer 1986).

The analysis of Attitude is based on the manual annotation and categorization of instances. In light of the degree of subjectivity which is involved in this process (Hunston 2004), we have carried out an inter-coder agreement test on a sample from the corpus. The test yielded a chance-corrected coefficient of k = 0,62 (Cohen 1960), which indicates a 'substantial' level of agreement and can be thus taken as a positive indicator of the reliability of identification and quantification of Attitude in the corpus.

The analysis of Engagement has been carried out using an automatic procedure for the quantification of 'markers' of Engagement. Engagement lends itself better than Attitude to software applications, as it is possible to identify in advance a circumscribed set of resources that can be searched for and quantified in the corpus. For our analysis we have assembled two collections of potential markers of Engagement, created adapting and integrating the lists of 'stance markers' provided in Biber and Finegan (1989).

The analysis highlights significant differences in the use of Appraisal in the two reports. BP deploys attitudinal language to strongly foreground its technical capabilities and expertise. Ikea displays affect, emphasizes improvements and hedges propositions more frequently then its counterpart. This different use of evaluation construes two very different institutional and corporate identities, which can be read and decoded in view of the specificities of the contexts in which the two companies operate and of the sustainability challenges they have to face in their daily operations.

References

[1] Biber, D. and Finegan, E. (1988). Adverbial stance types in English. Discourse Processes 11, 1-34.

[2] Biber, D., Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. Text 9, 93-124.

[3] Cohen, J. (1960). A coeffcient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46.

[4] Chafe, W. and Nichols, J. (eds.). (1986). Evidentiality: the Linguistic Coding of Epistemology. Norwood, N.J.: Ablex.

[5] Fairclough, N. (2003). Analysing Discourse. London and New York, Routledge.

[6] Hunston, S. and Thompson, G. (eds.). (2000). Evaluation in Text: Authorial Stance and the Construction of Discourse. Oxford, OUP.

[7] Hunston, S. (2004). Counting the uncountable: problems of identifying evaluation in a text and in a corpus. In Partington, M. and Haarman, L. (eds.), Corpora and Discourse. Peter Lang, 157-188.

[8] Hyland, K. (1996). Writing Without Conviction: Hedging in Science Research Articles. Applied Linguistics 17, 433-54.

[9] Macken-Horarik, M., and Martin. J.R., (eds.). (2003). Text 23. Special Issue. Negotiating Heteroglossia: Social Perspectives on Evaluation. Berlin and New York, Mouton de Gruyter.

[10] Malavasi, D. (2007). Lexical analysis of implicit promotional devices in Bank Annual Reports. Les Cahiers de lILCEA numro 9, 171-184.

[11] Malavasi, D. (2008). Banks Annual Reports: an overview of the linguistic means used to express evaluation. In Garzone, G. and P. Catenaccio (eds.), Language and Bias in Specialized Discourse. Milano, CUEM, 139-152.

[12] Martin, J. R. (1995). Reading Positions/Positioning Readers: JUDGEMENT in English. Prospect: a Journal of Australian TESOL 10 (2), 27-37.

[13] Martin, J.R. (2000). Beyond Exchange: APPRAISAL Systems in English. In Hunston, S. and Thompson, G. (eds.), Evaluation in Text. Oxford, Oxford University Press, 142-75.

[14] Martin, J. R. and White, P. R. R. (2005). The Language of Evaluation: Appraisal in English. New York and London: Palgrave.

[15] Palmer, F. R. (1986). Mood and Modality. Cambridge: Cambridge University Press.

[16] White, P. (2001). An introductory tour through appraisal theory. http://www.grammatics.com/appraisal/AppraisalGuide/Framed/Appraisal-Overview.htm

How to choose among synonyms: Corpus evidence for a “synonymy cline”

Gregory Garretson
Uppsala University

While it is generally agreed that perfect synonymy does not occur, there are many sets of words that are broadly substitutable for each other without an obvious change in meaning. How, then, do speakers choose a word from a set of synonyms? Corpus linguists know well that context is a powerful determinant of lexical selection. What can we say about the constraints that cause a speaker/writer in the act of communicating to select one synonym over another? In this talk, I report on a study that leads me to propose what I term a “synonymy cline”, in which a number of different factors contribute to ranking a set of synonyms in terms of their suitability for a given context.

The synonyms selected for the study are (the relevant senses of) the nouns “sort”, “kind”, and “type”. Evidence for considering these to be synonyms includes the excellent body of work on these “type nouns” by Davidse et al. (2008), De Smedt et al. (2007) and colleagues, which, despite thorough analysis of the syntactic and discourse properties of these words, never draws distinctions among them. Nevertheless, corpus data reveal that these words are not fully interchangeable. The only work thus far that has drawn an explicit distinction among them is Biber et al. (1999). The present study confirms Biber et al.’s findings and takes the analysis much further.

This study examines the differences between “sort”, “kind”, and “type” in the BNC, focusing in particular on their collocational patterns. It finds that “sort” shows a strong preference for spoken language, informal language, and collocation with frequent words, while “type” shows a strong preference for written language, formal language, and collocation with infrequent, more technical words. In every single analysis, “kind” occupies an intermediate position, painting a clear picture of a gradient range of acceptability, or “synonymy cline”.

Based on these findings, I present a set of heuristics that may serve as a first approximation to a model of the unconscious constraints operating when a speaker must make a choice among these words. The overall effect resembles a set of gradient constraints conspiring to make “sort” the preferred choice for certain situations, “type” the preferred choice for others, and “kind” an ideal compromise when there is conflict between the constraints. I suggest that the same mechanism may be in operation in other cases of synonymy in English and other languages.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. 1999. Longman grammar of spoken and written English. New York: Longman.

Davidse, K., Brems, L., & Smedt, L. D. 2008. Type noun uses in the English NP: A case of right to left layering. International Journal of Corpus Linguistics, 13(2), 139–168.

De Smedt, L., Brems, L., & Davidse, K. 2007. NP-internal functions and extended uses of the 'type' nouns Kind, Sort, and Type: Towards a comprehensive, corpus-based description. In R. Facchinetti (Ed.), Corpus linguistics 25 years on (pp. 225–255). Amsterdam: Rodopi.

'Well, however, I shall leave this bad city to-morrow': connecting sentences in 18th century novels

Victorina Gonzalez-Diaz
University of Liverpool

The history of English connectives (i.e. coordinators—conjuncts—subordinators; Quirk et al. 1985: 926) seems to have received relatively little scholarly attention (Lenker 2010: 2-3). By the side of studies on either specific connectors or semantic relations of connectivity (see Lenker 2010 and the references provided there), Kortmann (1997), Adamson (1998, 1999), Kohnen (2007) and Lenker (2010) can be considered the most systematic diachronic treatments of different connectives types.

In line with previous scholarship, this paper contributes to a better understanding of the history of English connectives. Like Lenker (2010), my study is corpus-based and pays special attention to conjuncts (e.g. *however, therefore*). However, in keeping with Adamson (1998) and Kohnen (2007), the main aim of the study is stylistic, i.e. to explore the ways in which connectives are exploited for literary purposes.

More specifically, my case-study concentrates on the novel of sensibility (1740-1790), which develops during a period of stylistic transition (from the 'perspicuous' Enlightenment to 'expressive' Romanticism; Abrams 1953, Adamson 1998). As such, its discourse has been said to represent a "constant" negotiation between reason (Enlightenment) and feeling (Romanticism) (Bray 2003: 92). However, no analysis has, to date, examined how that negotiation is *linguistically* manifested. To this aim, the use of connectives constitutes a suitable case-study: subordination is the unmarked connective option in the early eighteenth century (Enlightenment) because it makes clear the semantic relation established between the connected clauses (Adamson 1998: 634); whereas Romantic discourse favours the conveyance of strong feelings though coordinators and asyndesis.

Through a quantitative and qualitative analysis of connectives in grammatical treatises and selected sentimental novels (*Clarissa*, *Sidney Bidulph*, *Julia de Roubigné*, *Evelina*), this paper suggests

a) that sensibility is *not* a linguistically homogeneous tradition that moves towards coordinative linking strategies the closer it gets to the 19th century;
b) that conjuncts (the intermediate category in Quirk et al's (1985) cline of sentence connectivity, above) are exploited by writers as character- and class-based indicators, and, in this connection,
c) that Burney's *Evelina* stands out amongst contemporaneous novels in terms of frequency and stylistic variety of connective uses.

Taken as a whole, then, my paper resonates strongly with the main themes of the conference: it constitutes an example of the fast-developing 'corpus stylistics' trend within corpus studies and provides a contrastive analysis of grammar-based distinctions across time and authors.

Assessing data-driven methods in onomasiological investigations: exploring quantification in business discourse

Diane Goossens
Université catholique de Louvain

Expressing quantification is a key element in business discourse. Quantification may be expressed in many different ways, including numerals (*500*), words (*substantially*), parts of words (*multi*-billion) or even punctuation marks (50-60). This paper sets out (1) to explore, test and evaluate various data-driven methods (Rayson, 2008) that can be applied to uncover the linguistic devices used to express quantification in two corpora of business English, and (2) to analyse the results yielded by these methods. Both precise (e.g. *58.3%*) and imprecise quantification (or approximation, e.g. *about 600 employees*) are examined. Two corpora of 1 million words are investigated: a corpus of business news reporting, the Business English News corpus (BENews), and a corpus of academic publications on a number of business topics, the business subcorpus of the Louvain Corpus of Research Articles (LOCRA_Business) (Centre for English Corpus Linguistics, Université catholique de Louvain). Several data-driven methods which are tested in this paper involve the use of annotated corpora. BENews and LOCRA_Business have been annotated using *Wmatrix* (UCREL, Lancaster University), a web-based corpus processing environment giving access to the CLAWS7 part-of-speech tagger, a lemmatizer and the *UCREL Semantic Analysis System* (USAS) (see Rayson, 2003). The use of part-of-speech tagged corpora makes it possible to start from specific grammatical categories to explore ways of expressing quantification and the preferred patterns in which they occur. For example, all the instances of tags for numbers (e.g. *MC*) can be extracted from the corpora and analysed using Concord in WordSmith Tools (Scott, 2004). Lemmatised words can also be classified according to their word classes and frequencies in the corpora and scrutinized for items potentially referring to quantification. In addition, as the corpora have been semantically annotated, several tags referring to semantic fields relating to quantification (e.g. *N5* for 'quantities' and *A13.4* for 'degree: approximators') can be automatically retrieved and then examined. Finally, WordSmith Tools' keyword analysis is tested using a 1-million-word 'fiction' reference corpus to highlight potential quantification devices that are specific to the business genres under study. The purpose of this paper is not only to test and discuss the advantages and disadvantages of these different methods but also, on the basis of an analysis of the results, to evaluate how these methods might successfully complement each other within the framework of onomasiological studies such as the investigation of the notion of quantification.

References

Rayson, P. (2003) *Matrix: a statistical method and software tool for linguistic analysis through corpus comparison.* Unpublished PhD thesis, Lancaster University. Available from http://www.comp.lancs.ac.uk/~paul/public.html

Rayson, P. (2008) From key words to key semantic domains. *International Journal of Corpus linguistics,* 13 (4): 519-49.

Scott, M. (2004) *WordSmith Tools 4*. Oxford: Oxford University Press.

A corpus-based study of gender assignment in English loanwords

Anne-Line Graedler
Hedmark University College

English is a source of extensive lexical borrowing in present-day European languages. As part of the morphological integration process, borrowed English nouns in Norwegian are assigned one of three grammatical genders. Although gender assignment has been dealt with in previous studies of English loanwords in Norwegian (e.g. Graedler 1998; Johansson & Graedler 2002), it is only during the past decade or so that large and systematic corpora of Norwegian have been made accessible for research (see the web pages of The Text Laboratory), and thus made it possible to approach Onysko's ideal that "a comprehensive analysis of English influence [...] should be based on large corpora of written and spoken [... language] in formal and informal settings from a variety of media" (2007: 98).

Several relatively recent studies take a principled view of gender attribution in Norwegian as rule-based or dependent on inherent schemas (e.g. Trosterud 2001; Enger 2001; 2002; 2004; 2009; Halse 2004; Ragnhildstveit 2009; Kristoffersen & Simonsen 2009). The assignment of gender in borrowed English nouns may be regarded as natural test cases for hypotheses about productive gender assignment, and may thus contribute valuable information to the study of gender assignment principles in general, and with respect to loanwords in particular.

In addition to presenting data from an investigation of gender assignment, the paper will address some methodological issues, such as the suitability of corpora for loanword identification and extraction, and questions of size and representativeness.

References

Enger, H.-O. (2001). Genus i norsk bør granskes grundigere. *Norsk Lingvistisk Tidsskrift* 19, 163-183.

Enger, H.-O. (2002). Stundom er ein sigar berre ein sigar. *Maal og Minne* (2), 135-151.

Enger, H.-O. (2004). On the relation between gender and declension: a diachronic perspective from Norwegian. *Studies in Language* 28, 51-82.

Enger, H.-O. (2009). The role of core and non-core semantic rules in gender assignment. *Lingua* 119, 1281-1299.

Graedler, A.-L. (1998). *Morphological, semantic and functional aspects of English lexical borrowings in Norwegian.* Oslo: Scandinavian University Press.

Halse, G.E. (2004). Genustilordning i nynorsk: Ei datamaskinell etterprøving. MA thesis, University of Bergen.

Johansson, S. & Graedler, A.-L. (2002). *Rocka, hipt og snacksy: Om engelsk i norsk språk og samfunn.* Kristiansand: Høyskoleforlaget.

Kristoffersen, K.E. & Simonsen, H. G. (2009). Tilegnelse av genus hos norske, danske og islandske barn. Presentation of a research project, Cognitive summer seminar, Hamar, June 2009.

Onysko, A. (2007). *Anglicisms in German: Borrowing, Lexical Productivity, and Written Codeswitching.* Berlin, New York: De Gruyter.

Ragnhildstveit, S. (2009). Genustildeling og morsmålstransfer i norsk mellomspråk. En korpusbasert studie. MA thesis, University of Bergen.

Trosterud, T. (2001). Genus i norsk er regelstyrt. *Norsk Lingvistisk Tidsskrift* 19, 29-58

Developing and analyzing the New Zealand part of the Engineering Lecture Corpus

Lynn Elaine Grant
Auckland University of Technology

Universities in three different countries – England, Malaysia and New Zealand – have been collaborating to develop the Engineering Lecture Corpus (ELC*). Each university is videoing and transcribing twenty hours of Engineering lectures. These lectures are being marked-up with the Oxygen XML software. Pragmatic functions being marked-up include: greetings, prayers, housekeeping, preview content, review content, defining terms, reference to future profession, personal narratives, and humour (including teasing, self-recovery, self-denigration, black humour, disparagement of out-group member, register and word play, and mock threat). There are both similarities and differences in the use and frequency of these functions among the three countries. This paper will focus on the development and content of the New Zealand portion of the ELC. Initially, there were twenty hours of Engineering lectures from five different branches of Engineering videoed and transcribed, but later ten additional hours were added from one branch. Differences could be seen in the use or lack of use of functions in the New Zealand Engineering lecturers. For example, unlike in the other two universities, New Zealand lecturers do not begin their lectures with a greeting, nor do the lecturers use the same type of humour. Differences can also be seen in the five different branches of Engineering. Thompson (2003) found that lecturers used different 'markers' to indicate transition from one (sub-) topic to another, by making reference to the content of the talk, to the talk itself or an interpersonal reference to the speaker or the audience. Examples of these will be identified in the New Zealand engineering lectures. And both Thompson (2003) and Flowerdew and Miller (1997) noted that materials for preparing EAL (English as an additional language) students for academic lectures are based on non-authentic lecture examples. The development of freely available academic corpora like MICASE (Michigan Corpus of Academic Spoken English) and BASE (British Academic Spoken English), plus more specialized corpora like the ELC, should help students by providing examples of authentic lectures in different fields.

*ELC is a project under the PMI2 project initially funded by the British Council

References

Flowerdew, J. & Miller, L. (1997). The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes, 16*(1), 27-46.

Thompson, S.E. (2003). Text-structuring metadiscourse, intonation and the signalling of organisation in academic lectures. *English for Academic Purposes,* 2, 5-20.

Disciplinary differences in small-group interactions: corpus perspectives on turn-taking in university seminars

Nicholas Groom, Oliver Mason
University of Birmingham

The fundamental aim of seminars and other forms of small-group interaction in higher education is to get students to talk, and the underlying assumption shared by educational theorists and university teachers alike is that the more the students talk, the more successful the seminar is. But what does 'more talk' mean? The number of words spoken, the number of turns taken, or the average length of turn? In this paper we report on a study of the British Academic Spoken English Corpus (BASE), in which we investigate student and teacher contributions to seminars according to each of these three measures. Our main quantitative finding is that different knowledge domains perform better according to different measures. Specifically, our analysis shows that students talk the most in seminars in the social sciences and humanities if we define talk in terms of total words spoken; students in physical sciences talk the most if we quantify talk in terms of number of turns; and students in life sciences talk the most if we measure talk in terms of average turn length. We then argue against the idea that any one of these measures might be inherently better or more desirable than the others. Drawing on qualitative data from the BASE seminars subcorpus, we argue instead that each of these different versions of 'talking more' carries with it a different set of affordances, each of which is more or less well attuned to the particular epistemologies and pedagogic goals of different academic disciplines. We conclude by considering the implications of our analysis for staff development and training programmes in higher education.

Fluency in native and non-native English speech: Theory, description, implications

Sandra Götz
Justus Liebig University, Giessen

Fluency is a widely used notion when speaking about – and assessing – both native and non-native speech. Previous research on fluency has shown, however, that describing its linguistic substance "with a degree of consensus is notoriously difficult" (Hasselgren 2002: 147), because various definitions of the concept of fluency co-exist and fluency is an epiphenomenon to which many individual (and interrelated) factors contribute. Most of the fluency-related research so far has focused on only one of the following aspects:

(1) temporal variables in speech production, such as length of runs, pause ratio, speech rate, etc. (e.g. Lennon 1990, Chambers 1997, Cucchiarini et al. 2002, Gut 2009);
(2) the use of prefabricated units and formulaic sequences (e.g. Pawley & Syder 1983, Wray 2002, Erman 2007);
(3) certain performance phenomena which serve as communication management strategies to cope with the planning pressure in online speech production, such as self-repairs, hesitation phenomena, the use of discourse markers, etc. (e.g. Biber et al. 1999, Hasselgren 2002, Rühlemann 2006).

The present paper combines these three approaches to productive fluency and defines 'fluencemes', i.e. functional categories abstracted from these features of spoken language that are all relevant for establishing a speaker's fluency, e.g. the range and frequency of discourse markers.

Based on this taxonomy, firstly, a comprehensive contrastive analysis of the individual fluencemes of the learner data of the 86,000-word German learner corpus LINDSEI-GE and the comparable 119,000-word native speaker corpus LOCNEC will be presented which identifies areas in which advanced German learners of English still deviate strongly from the native target norm (e.g. in their range and distribution of formulaic sequences) and others in which some of the speakers have already approximated to the target norm (e.g. in their use of filled pauses).

Secondly, I will present the findings of an analysis that focused on how the individual fluencemes relate to each other. This analysis revealed that a nativelike fluency performance can be characterized by certain patterns: There are fluencemes that appear in each speakers' output (e.g. a similar proportion of formulaic sequences), whereas other fluencemes can be considered as exchangeable '*allo*-fluencemes' (e.g. speakers either use a high proportion of discourse markers *or* smallwords *or* filled pauses). For the learner data, however, no such fluency clusters become that clearly visible and not all fluencemes are represented to the same extent in the learner output, which suggests that the learners have not internalized a nativelike variability in their fluenceme usage.

In a last part, I will discuss some language-pedagogical implications derived from these findings.

References

Chambers, F. (1997): "What do we mean by fluency?", *System*, 25 (4), 535-544.

Cucchiarini, C., H. Strik & L. Boves (2002): "Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech", *Journal of the Acoustical Society of America*, 111 (6), 2862-2873.

Erman, B. (2007): "Cognitive processes as evidence of the idiom principle", *International Journal of Corpus Linguistics* 12 (1), 25-53.

Gut, U. (2009): *Non-native speech. A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Frankfurt am Main: Peter Lang.

Hasselgren, A. (2002): "Learner Corpora and language testing – Smallwords as markers of learner fluency". In S. Granger, J. Hung & S. Petch-Tyson (eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins. 143-173.

Lennon, P. (1990): "Investigating fluency in EFL: a quantitative approach", *Language*
Learning, 40 (3), 387-417.

Pawley, A. & F. H. Syder (1983): "Two puzzles for linguistic theory: Native-like selection and native-like fluency", *Language and communication*, ed. J. Richards & R. Schmidt. London: Longman. 191-226.

Rühlemann, C. (2006): "Coming to terms with conversational grammar: 'Dislocation' and 'dysfluency'", *International Journal of Corpus Linguistics* 11(4), 385-409.

Wray, A. (2002), *Formulaic language and the lexicon*, Cambridge: Cambridge University Press.

The conceptual convergence of functional-cognitive theory and neo-Firthian linguistics

Andrew Hardie, Tony McEnery
Lancaster University

A recent trend in corpus linguistic research has been the application of corpus methods to questions within the framework of functional-cognitive linguistic theory (see Gries 2006, Gilquin and Gries 2009, Arppe et al. 2010). Notable is the development by Stefanowitsch and Gries (2003) of *collostructional* analysis, an approach to lexical-grammatical co-occurrence patterns which brings statistical approaches to collocation together with a cognitive-theoretical framework, namely, Construction Grammar (Goldberg 1995, Croft 2001).

The joining of corpus methods and cognitive theory has met with objections from certain researchers within the *neo-Firthian* school of corpus linguistics, that is, the tradition drawing inspiration from John Sinclair. These objections, expressed in sometimes highly vituperous terms (Louw 2010, Teubert 2005, 2010), are generally in accordance with the core neo-Firthian precept of rejecting any theoretical construct arrived at separately from the corpus. In debates on this issue, an often tacit assumption is that the neo-Firthian approach and corpus-based functional-cognitive linguistics are actually divergent. We wish to argue that functional-cognitive theories (like Construction Grammar), and advanced neo-Firthian theories such as Lexical Priming (Hoey 2005) and Pattern Grammar (Hunston and Francis 1999), are, rather, profoundly convergent.

This convergence has been noted by researchers in other fields of linguistics; for instance, Ellis (2002), in a review of frequency effects on acquisition, explicitly espouses both Construction Grammar and Sinclair's Idiom Principle as explanations of these psycholinguistic phenomena. However, the convergence is equally apparent on close examination of the core concepts on each side of the comparison, which we will demonstrate in two ways. First, we will examine lexis/grammar co-selection according to Construction Grammar (i.e. collostruction) and according to neo-Firthian theory (i.e. *colligation* in Hoey's terms, *patterns* in Hunston and Francis'), and show these perspectives to be more directly equivalent than they are typically considered. Second, we will examine the descriptive/ theoretical apparatus of Pattern Grammar and Construction Grammar and show that the distinctions assumed or argued to exist between them cannot be maintained. Patterns or abstract colligations *are* constructions just as constructions *are* patterns. Differences persist in terminological and methodological preferences – notably regarding the degree of abstraction away from concrete instances of usage deemed appropriate *at the outset* of an analysis.

Finally, we will argue that this convergence of functional-cognitive and neo-Firthian theories suggests that the core matter of their convergence is, to some approximation, 'the truth' about language. We will suggest a tentative formulation of the theory of language that this implies.

References

Arppe, A, Gilquin, G, Glynn, D, Hilpert, M and Zeschel, A (2010) Cognitive Corpus Linguistics: five points of debate on current theory and methodology. *Corpora* 5(1): 1-27.

Croft, W. (2001) *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Ellis, N. C. (2002). Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24: 143-188.

Gilquin, Gaëtanelle and Stefan Th. Gries (2009), Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1): 1–26.

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

Gries, S. Th. (2006) Introduction. In Stefanowitsch, A and Gries, S. Th. (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*. Berlin & New York: Mouton de Gruyter, pp. 1-17.

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

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

Louw, W.E. (2010) The painting is where the paint is not: Reflections on the 'Bootcamp Debate'. *International Journal of Corpus Linguistics* 15(3): 344-53.

Stefanowitsch, A and Gries, S. Th. (2003) Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209-43.

Teubert, W. (2005) My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1): 1-13.

Teubert, W. (2010) Our brave new world? *International Journal of Corpus Linguistics* 15(3): 354-8.

Quotative Use: A cross-variety comparison

Nicole Hoehn
University of Basel

This paper presents the results of a study comparing the quotative systems of Jamaican English, Irish English and Canadian English. In the last thirty years, exciting changes have taken place in the field of quotatives in that new verbs of quotation such as *be like* and *go* have emerged in varieties of English. Up to now, research on these new quotatives has only covered certain varieties of English. The studies published so far mainly focus on American English (e.g. Barbieri 2007), English English (e.g. Buchstaller 2006), Canadian English (e.g. Tagliamonte and D'Arcy 2007), Scottish English (e.g. Macaulay 2001), New Zealand English (e.g. D'Arcy 2010) and Australian English (e.g. Winter 2002). The lack of studies on quotatives in Jamaican English can be explained by the fact that research on the Jamaican speech community has traditionally paid attention to the basilectal end of the (post-) creole continuum as it was assumed that the acrolectal end was very similar to Standard English in Britain, the standard variety of the former colonizers. Recently, however, interest in the acrolect has increased. Sand (1999) and Mair (2002) point out that educated Jamaican English tends to move away from the inherited British norm and that "Jamaican Standard English is only now emerging" (Mair 2002: 31). My aim is to attempt to fill the gap in research on verbs of quotation in Jamaican and Irish English by investigating their use on the basis of data from the spoken parts of the Jamaican and Irish components of the *International Corpus of English* (ICE). Using a variationist approach, the paper explores to what extent the use of quotatives is affected by social variables such as speaker sex as well as by linguistic variables such as grammatical person of the quotative, tense of the quotative and content of the quote. To this end, all quotatives and zero quotatives have been extracted from the text category of private conversations, a process resulting in a total of more than four hundred tokens in the Jamaican dataset and more than one thousand tokens in the Irish dataset. The paper will show which social and linguistic factors condition the probability of the most frequently used (innovative) quotatives in Jamaican and Irish English. These findings will be compared with the use of quotatives in ICE-Canada. The three ICE corpora share a common design and roughly the same sampling period. Thus, they lend themselves to a comparison.

Canonical tag questions in Asian Englishes – A study in variational pragmatics

Sebastian Hoffmann, University of Trier
Joybrato Mukherjee, Justus Liebig University Giessen

Canonical tag questions (TQs) such as shown in (1) to (4) are a very conspicuous feature of spoken language, and their form (e.g. the polarity of anchor and question tag) as well as their pragmatic functions (e.g. seeking confirmation or facilitating the flow of interaction) have received ample coverage in the literature (e.g. Holmes 1983; Algeo 1988, 1990; Stenström 1994). They are particularly common in British English, where their frequency of use was shown to be nine times higher than in comparable data from American English (Tottie & Hoffmann 2006).

	Anchor	Question Tag	Polarity
(1)	You turned it round then	didn't you? (ICE-GB)	positive-negative
(2)	You you're not Cantonese	are you? (ICE-SIN)	negative-positive
(3)	Her surname is Nagouda	is it? (ICE-IND)	positive-positive
(4)	It is not on cats	isn't it? (ICE-SIN)	negative-negative

It has been shown in various studies that TQs are a vibrant and interesting area of variation in institutionalised second-language varieties of English (e.g. Sahgal & Agnihotri 1985 for Indian English, Cheng & Warren 2001 for Hong Kong English; Wong 2008 for Singapore English). Previous research, however, has tended to focus on individual postcolonial Englishes and has placed special emphasis on the use of invariant question tags (e.g. *no?*). A comprehensive comparison of the forms, functions and frequencies of canonical TQs across postcolonial Englishes is still missing. Our study aims to fill this gap by investigating TQ usage in the Hong Kong, Indian and Singapore components of the International Corpus of English (ICE); the British component ICE-GB will be used as the baseline variety, representing the present-day status of the historical input variety for all three postcolonial Englishes in question.

Our choice of varieties is primarily informed by the fact that they represent different stages in the development of postcolonial Englishes according to the evolutionary model of variety-formation suggested by Schneider (2007), with Hong English being the least advanced and Singapore English being the most advanced variety. Recent studies have revealed that there are clear correlations between the evolutionary stage of the three Asian Englishes on the one hand and the degree of structural nativisation at the lexicogrammatical level (e.g. with regard to collostructions) on the other (Mukherjee & Gries 2009; Gries & Mukherjee 2010). An investigation of canonical TQs in these three varieties will allow us to assess whether similar correlations between structural developments at the level of discourse pragmatics and the process of structural nativisation of postcolonial Englishes can be identified.

In general, our findings show that the process of structural nativisation also manifests itself in the changing use of canonical TQs, e.g. with regard to their increasing use as invariant TQs. From a theoretical perspective, our observations can be viewed as a contribution to a variational-pragmatic approach (e.g. Schneider & Barron 2008) to postcolonial Englishes.

References

Algeo, J. (1988): "The tag question in British English: It's different i'n'it? *English World-Wide* 9(2),171-91.

Algeo. J. (1990): "*It's a myth, innit?* Politeness and the English tag question." In: Ricks, C. & L. Michaels (eds.) *The State of the Language*. Berkeley: University of California Press, 443-50.

Cheng, W. & M. Warren (2001): "'She knows about Hong Kong than you do isn't it?': Tags in Hong Kong conversational English." *Journal of Pragmatics* 33(9), 1419-1439.

Holmes, J. (1983): "The functions of tag questions." *English Language Research Journal* 3, 40-65.

Gries, S.T. & J. Mukherjee (2010): "Lexical gravity across varieties of English: An ICE-based study of *n*-grams in Asian Englishes", *International Journal of Corpus Linguistics* 15(4), 520-548.

Mukherjee, J. & S.T. Gries (2009): "Collostructional nativisation in New Englishes: verb-construction associations in the International Corpus of English", *English World-Wide* 30(1), 27-51.

Sahgal, A. & R.K. Agnihotri (1985): "Syntax – the common bond: Acceptability of syntactic deviances in Indian English", *English World-Wide* 6(1), 117-129.

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

Schneider, K.P. & A. Barron (2008): *Variational Pragmatics: A focus on regional varieties in pluricentric languages.* Amsterdam: Benjamins.

Stenström, A.-B. (1994): *An Introduction to Spoken Interaction*. London: Longman.

Tottie, G. and S. Hoffmann. (2006): "Tag questions in British and American English." *Journal of English Linguistics* 34(4), 283-311.

Wong, J. (2008): "Anglo English and Singapore English tags: Their meaning and cultural significance", *Pragmatics and Cognition* 16(1), 88-117.

Factors conditioning the choice of relativizers in 18th and 19th century English – A diachronic study based on the Old Bailey Corpus

Magnus Huber
Justus-Liebig-Universitaet Giessen

The *Old Bailey Corpus* (*OBC*) is based on the *Proceedings of the Old Bailey*, London's central criminal court. They were published from 1674 to 1913 and constitute a large body of texts from the beginning of Present Day English (over 200,000 trials, ca. 134 million words), its verbatim passages being arguably as near as we can get to the spoken word of the period. The material offers the rare opportunity of analyzing everyday language in a period that has been neglected both with regard to the compilation of primary linguistic data and the description of the structure, variability, and change of English. The *OBC* identifies about 114 million words as direct speech from the 1720s onwards, and about a third of this material is being marked-up for sociolinguistic (sex, profession, age, residence of speaker, role in the court-room) and for textual variables like the shorthand scribe and publisher of individual *Proceedings*.

This paper will investigate relative clauses in 18th and 19th century spoken English as documented in the *Proceedings of the Old Bailey*. Although this phenomenon is by no means understudied, many publications on the history of the relative clause remain impressionistic and/or are based on written rather than spoken language. The *OBC*, because of its sheer size, the time span covered, and the available sociobiographical speaker and textual information, is an ideal corpus for a fine-tuned, quantitative-variational study of the relative clause at the beginning of Present Day English.

The paper will consider the influence of several independent variables on the choice of the main relativizers *who/m/se*, *which*, *that* and zero. This will include linguistic factors like the animacy of the antecedent, the syntactic role of the relativizer, the voice of the relative clause as well as extra-linguistic factors like speaker gender and the scribe/publisher of the *Proceedings*, who may have imposed their house style on the material. At the beginning of the period considered here, *that* and zero were still felt to be extremely colloquial, at least in writing (Görlach 2001: 126-127), and it will be interesting to see in what way the variation between *that*/zero and the pronouns *who(m)* and *which* develops over the two centuries and what changes in the determining variables we can observe. Other areas to be investigated are the variation between *who* and *which* for human antecedents and between *who* and *whom* for object relative clauses.

Reference

Görlach, Manfred. 2001. *Eigteenth-century English*. Heidelberg: C. Winter.

Negation in English and in Spanish: A corpus-based diagnosis

Marlén Izquierdo, Rosa Rabadán
University of León

This contribution is framed within the ACTRES project (http://actres.unileon.es/) and explores the distribution and use of the main negative resources in English and in Spanish- namely negative operators, negative existentials and affixal negation. It aims to verify whether this distribution is reflected in translated Spanish or if other options are favoured. The study is placed in the context of contrastive research and is carried out through a 4-tier procedure: i) selection, ii) description and juxtaposition, iii) contrast, and iv) verification of 'target language fit' (Chesterman 1998).

The study uses monolingual corpora, BoE (http://wordbanks.harpercollins.co.uk/auth/?module=login) and CREA (http://corpus.rae.es/creanet.html) and parallel corpus P-ACTRES together with technical (statistics) and evaluative tools (informants). The role of the large monolingual corpora is to provide empirical data concerning negation in English and in Spanish, whereas P-ACTRES (http://actres.unileon.es/corpussearch/) contributes empirical information about translation behaviour concerning negative items. In turn, CREA acts as a control corpus for non-translated Spanish.

The data so analyzed provide information about: i) the English and Spanish resources that convey negative meanings and their relative centrality, ii) the translational solutions to bridge the cross-linguistic differences, and iii) the disparities between negative uses in translated and non-translated Spanish.

Empirical data suggests a marked overuse of affixal negation in translated Spanish as opposed to regular native usage. It also reveals the promotion of formally different Spanish expressive strategies to convey delicately nuanced negative English contexts (e.g. double negation as hedging, as in *not unusual, not blameless*). The former seems to affect the scope of negation and be the cause of unwanted semantic and pragmatic shifts in the translations; the latter sheds light on resources frequently overlooked that, however, seem to serve these contexts well.

These results can be further systematized as prescriptively descriptive guidelines which might be useful in applied areas (Rabadán 2010).

References

Chesterman, A. 1998. *Contrastive Functional Analysis*. Amsterdam/Philadelphia: John Benjamins.

Hernández Paricio, F. 1985. *Aspectos de la negación en español*. Universidad de León. Contextos 3.

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

RAE. 2009. *Nueva Gramática de la Lengua Española. Sintaxis II*. Madrid: Espasa Libros (Asociación de Academias de la Lengua Española). 3633-3715.

The development of comment clauses: accounting for current change

Gunther Kaltenböck
University of Vienna

Although the exact origin and historical pathway of comment clauses such as *I think, I suppose, I guess* are a matter of some discussion (e.g. Thompson & Mulac 1991, Brinton 2008, Fischer 2007), it is generally accepted that they have grammaticalized from fully lexical clauses into epistemic markers (e.g. Brinton 2008, Van Bogaert 2009). This paper explores to what extent in recent decades there is evidence of further grammaticalization of comment clauses, using data from the *Diachronic Corpus of Present-Day Spoken English* and the *Corpus of Historical American English.*

Taking into account a range of different parameters, such as overall frequency and positional distribution, use of the *that*-complementizer, semantic-pragmatic scope over the host construction, collocation patterns, and pragmatic functions, the study shows that a comment clause such as *I think* is increasingly used not so much as an epistemic marker but as a textual/interactional device (cf. Kärkkäinen 2003, Kaltenböck 2010). This erosion of the epistemic function of *I think* can be linked to an increased use of variant forms in recent years, notably *I'm thinking, I just think, I'm guessing*, which seem to be recruited as epistemic markers to compensate for the fading modal meaning of *I think*.

To account for the development of comment clauses, the paper finally argues for a Construction Grammar approach which places comment clauses in a larger constructional network with taxonomic links to related constructions, viz. the matrix-object clause (transitive) schema and the sentence adverbial schema. This view can account for the advance of *I think* from an epistemic to a general pragmatic marker as well as the use and retention of the *that*-complementizer.

References

Brinton, Laurel J. 2008. The comment clause in English. Syntactic origins and pragmatic development. Cambridge: C.U.P.

Fischer, Olga. 2007. Morphosyntactic change. Functional and formal perspectives. Oxford: O.U.P.

Kaltenböck, Gunther. 2010. Pragmatic functions of parenthetical *I think*. In Gunther Kaltenböck, Wiltrud Mihatsch, Stefan Schneider (eds.). *New approaches to hedging*. Bingley: Emerald, 243-272.

Kärkkäinen, Elise. 2003. Epistemic stance in English conversation: A description of its interactional functions, with a focus on I think. Amsterdam: Benjamins.

Thompson, Sandra A.; Mulac, Anthony. 1991. "The discourse conditions for the use of the complementizer *that* in conversational English". *Journal of Pragmatics* 15: 237-251.

Van Bogaert, Julie. 2009. The grammar of complement-taking mental predicate constructions in present-day British English. PhD dissertation, University of Gent.

Birmingham Blog Corpus: a new diachronic corpus of blog posts

Andrew Kehoe, Matt Gee
Birmingham City University

In recent years, there have been two main approaches taken to the 'web as corpus'. One has been to treat the web itself as a vast corpus, searchable through commercial search engines or specialist tools such as WebCorp. The other approach has been to use the web as a large archive from which texts can be selected for inclusion in structured corpora. The Corpus of Contemporary American English (COCA) is one example of the latter approach. Although COCA contains texts downloaded from the web, the corpus does not contain new web-specific textual varieties such as blogs, message boards or other computer-mediated communication; the written genres in the corpus are similar to those found in standard corpora: fiction, popular magazines, newspapers, academic journals. (This is for good reason: maintaining the genre balance across the years 1990-2010 would be impossible if web-specific genres were included.)

In fact, there are relatively few large-scale corpora of 'web-native' (sub-) genres. There have been many ethnographic and socio-linguistic analyses of blog data, but most use small manually-collected datasets. This paper presents work on the new Birmingham Blog Corpus (BBC): a 100 million word, diachronically-structured corpus of blog posts and reader comments. The corpus is part-of-speech tagged, annotated for textual domain, and publicly searchable through the WebCorp Linguist's Search Engine interface.

The first part of the paper describes the steps involved in building BBC, including a discussion of the sources chosen for blog data, the 'seeding' techniques used, and the corpus design decisions made. We then examine the characteristics of the blog genre, using metrics such as POS frequencies, sentence, paragraph and document length, and HTML layout features. This is achieved by comparing the blog corpus with a 10 billion word general web corpus. We use the term 'genre' in a Swalesian sense: "a class of communicative events ... [sharing] some set of communicative purposes" with similar structure and stylistic features.

In the second part of the paper, we move towards the content or 'aboutness' level. Many blog posts are classified according to topic, by the author or by third-party catalogues and syndicators such as Technorati and Google Blog Search. We carry out a lexical analysis, using a 'keywords' approach, to determine how well such topic categorisation reflects textual reality. We also consider the reader perspective by analysing the lexis of blog comments and the topic-related labels, or 'tags', assigned to blog posts through sites such as Delicious.

Standard English and ICE-Ireland

John M. Kirk
Queen's University Belfast

In this paper I attempt a re-appraisal of the notion of 'standard' English as an empirical standard. By using the spoken component of the ICE-Ireland corpus as data, my approach is to describe what is to be found in each particular speech situation, paying particular regard to the North-South zones, medium, and register, which were the main factors in the selection of the corpus's data. I demonstrate what speakers of the Irish version of international standard English say; in so doing, I am able to demonstrate that very variability within Standard English which enables flexibility in language use – a flexibility which would be impossible if the standard were an idealised, invariant, static, change-and-variation-inhibiting form. People presume themselves to be speaking standard(ised) language insofar as they are attempting to conform to idealisations within an underlying belief-system about a fixed, invariant norm. And indeed ICE-Ireland adduces in abundance all the major syntactic structures of English which are shared universally and world-wide; but it also has the presence of a broad range of supposedly non-standard morpho-syntactic features, which will be exemplified. On the basis of the ICE-Ireland evidence, I am able to show that the situationally-based, ideologically-driven standard English of ICE-Ireland is, in form, structure and function, not fully standardised. In short, ICE-Ireland presents us with an empirical standard.

A multivariate approach to the word-class distribution in L2 spoken English

Yuichiro Kobayashi
University of Osaka

This study is one of a series which aims to determine Japanese EFL learners' developmental indices by describing and analyzing their performance as observed in learner corpus data. In this study, multivariate analyses are conducted on a spoken corpus of Japanese learners of English, using word-class frequencies as variables, in attempt to examine complex interrelationships between word-classes, those between different proficiency levels, and those between word-classes and levels.

This study draws on the NICT JLE Corpus, a corpus of Japanese EFL learners' oral interview transcripts. It consists of 1,200 examinees who have taken Standard Speaking Test (SST), and the test has 9 different levels to assess speaking proficiency (Izumi *et al.*, 2004). The corpus was tagged with the CLAWS, using the C7 tagset (Garside & Smith, 1997).

The approach adopted in this study has three characteristics. First of all, it is corpus-based. As previous studies of language acquisition have been restricted to relatively small amounts of data, research using larger data sets may lead to significant advances in the understanding of language acquisition (Biber, Conrad, & Reppen, 1998). Second, it focuses on word-class tags. Representing a word by its word-class will prevent a thematic difference between texts from eclipsing subtler stylistic differences (Tabata, 2002). Third, it is based on multivariate analysis. This study employs two multivariate techniques for data-reduction, correspondence analysis and cluster analysis.

Using the frequencies of 135 word-class tags, I will conduct a correspondence analysis in order to explore complex interrelationships between 135 word-classes and 9 proficiency levels. As for the result of the correspondence analysis, I will focus on the most powerful dimension, which account for 81.29 % of total variation in the data matrix. The most prominent feature of the result is that the dimension makes a contrast between novice and advanced learners. While 4 sub-corpora of the NICT JLE Corpus (Lv.1-Lv.4) have negative scores for the dimension, 5 sub-corpora (Lv.5-Lv.9) have positive scores. As for the distribution of word-class tags in the dimension, noun-related tags are characteristic of novice learners, and verb-related tags are characteristic of advanced learners. The result of the cluster analysis also shows that the frequencies of nouns and verbs make a similar contrast. Gentner (1982) points out that across many of the world's languages children initially learn nouns more readily than verbs. It is very interesting that the tendency is observed in L2 development.

References

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In

Garside, R., Leech, G., & McEnery, A. (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102-121). Longman, London.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In Kuczaj, S. A. (Ed.), *Language, thought, and culture* (pp, 301-334). Hillsdale: Lawrence Erlbaum.

Izumi, E., Uchimoto, K. & Isahara, H. (2004). *A speaking corpus of 1,200 Japanese learners of English.* Tokyo: ALC Press Inc.

Tabata, T. (2002). Investigating stylistic variation in Dickens through correspondence analysis of word-class distribution. In Saito, T., Nakamura, J., & Yamazaki, S. (Eds.), *English corpus linguistics in Japan* (pp. 165-182). Amsterdam: Rodopi.

Network corpora and corpus networks

Thomas Kohnen, Tanja Ruetten
University of Cologne

Historical corpus projects usually take the notion of *genre* as a vantage point, but there are various approaches to *genre* which seem to be involved in individual projects. Most simply, there are corpora that focus on single genres (for example, letters, as in the *Corpus of Early English Correspondence*). Then, there are corpora that contain several genres which are linked by a common medium (spoken language, *A Corpus of English Dialogues*), by a common publication type (pamphlets, *The Lampeter Corpus*; newspapers, *Zurich English Newspaper Corpus*) or by a common domain (medical discourse, *Corpus of Early English Medical Writing*; religious discourse, *Corpus of English Religious Prose*).

The texts and (sub-)genres of such corpora, especially those that are based on a common domain, share specific discourse functions, aims, topics and participants. Consequently, they may be seen as networks that involve "hierarchies" (e.g. in terms of prestige or relevance for the discourse community), logical "chains" and "sets" of genres (Swales 2004: 23; see also Swales 1990).

Network structures of these kinds and their implications for corpus compilation and linguistic analysis have so far not been systematically investigated. Not much research has been devoted either to the question in how far such 'network corpora' offer links that can help to connect different (domain-based) corpora to form 'corpus networks'.

In our presentation, we will first give an overview of the network structure of the *Corpus of English Religious Prose* (*COERP*), focussing on the threefold distinction between core, peripheral and associated genres. We will illustrate the analytical power of this distinction by presenting a pilot study of performativity in the various genre sets. We will show that the hierarchies, chains and sets that form the religious genre network have a strong bearing on the evolution of individual genres, and that text- and genre-based language change can best be understood as a coordinated process of the genres in the network (cf. Rütten in press, Kohnen 2010).

In the second part we will explore the possible links that may connect *COERP* to other genre- or domain-based corpora. Possible factors to be reviewed will be general text functions (for example, exposition, narration, exhortation), the subordinate/superordinate and central/peripheral position of genres in the domain, the position of authors and discourse communities.

In all, our paper will thus explore the value of considering corpora as networks in two complementary views.

References

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

Corpus of Early English Medical Texts. Middle English Component: 2005. Taavitsainen Irma, Päivi Pahta and Martti Mäkinen (eds.). CD-ROM. Amsterdam: John Benjamins.

Corpus of English Dialogues 1560-1760. 2006. Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University).

Corpus of English Religious Prose. Currently being compiled under the supervision of Thomas Kohnen (University of Cologne). See www.helsinki.fi/varieng/CoRD/corpora/COERP for details on design and text coverage.

Kohnen, Thomas. 2010. "Religious Discourse". In: Andreas H. Jucker and Irma Taavitsainen (eds.) *Historical Pragmatics*. Berlin: Mouton de Gruyter. 523-547.

Lampeter Corpus of Early Modern English Tracts. 1999. Compiled by Josef Schmied, Claudia Claridge, and Rainer Siemund. (In: ICAME Collection of English Language Corpora (CD-ROM), Second Edition, eds. Knut Hofland, Anne Lindebjerg, Jørn Thunestvedt, The HIT Centre, University of Bergen, Norway.

Rütten, Tanja (in press). *How to Do Things with Texts. Patterns of Instruction in Religious Discourse 1350-1700*. Bern: Peter Lang.

Swales, John. 1990. *Genre Analysis. English for academic and research settings*. Cambridge: Cambridge University Press.

Swales, John M. 2004. *Research Genres. Exploration and Applications*. Cambridge: Cambridge University Press.

Zurich English Newspaper Corpus. 2006. Lehmann, Hans Martin, Caren auf dem Keller, Beni Ruef. 2006. Zen Corpus 1.0. In: Roberta Facchinetti and Matti Rissanen (eds.), Corpus-based Studies of Diachronic English. 135-155. (Linguistic Insights 31.) Bern: Peter Lang.

"The amazing thing about this love story": On the use of generic noun *thing* as a function word in English and abstract *lo*-nominalizations in Spanish

Belén Labrador
University of León

This article aims at exploring the relationships between generic noun *thing* as a function word in English and Spanish neuter article *lo* as a nominalizer of adjectives on the basis of a perceived similarity in their semantic and pragmatic functions. Due to the fact that both structures occur in English and in Spanish but they differ greatly in their uses, this is an area of discrepancy in English-Spanish contrastive grammar which causes problems to Spanish EFL students and sometimes results in lack of idiomaticity in the target texts produced by Spanish translation students. Although de-adjectival nominalization with a definite article is possible in both languages, there are restrictions for its use in the case of English as compared with Spanish – English articles and adjectives are invariable in form whereas the morphological nature of Spanish articles and adjectives allows for number and gender distinction, which enables neuter *lo* to become specialized in abstract nominalizations. On the other hand, overuse of *cosa* as a literal translation of *thing* renders into grammatical but unnatural, even unacceptable, Spanish. Two monolingual corpora – Collins Wordbanks Online, for English: http://www.collinslanguage.com/wordbanks/and CREA, for Spanish: http://corpus.rae.es/creanet.html – and a parallel corpus – P-ACTRES: http://actres.unileon.es/inicio.php?elementoID=12 (composed of original English texts and their corresponding translations into Spanish) have been used for the purpose of this study. The results show that both *thing* as a function word and *lo* as a nominalizer are highly productive grammatical resources – co-occurrences with a wide range of different adjectives have been found; however, they both tend to concentrate on a series of adjectives, which vary to a certain extent in the two languages involved. A number of other similar expressions reflecting the abstract quality usually expressed by the adjective are revealed by a bidirectional analysis, i.e. a) translations of the English pattern with *thing* as a function word into Spanish and b) source expressions of the Spanish occurrences of *lo*-nominalizations.

Writing in tables and lists: Exploring multimodal undergraduate writing through keyword searches

Maria E Leedham
The Open University

Chinese people now comprise the 'largest single overseas student group in the UK' with more than 85,000 registered at UK educational institutions in 2009 (British Council, 2010). While Chinese students' academic writing in English has been explored through corpora comprising short argumentative essays (e.g. Mayor et al, 2007; Chuang and Nesi, 2006), or postgraduate theses (e.g. Hyland, 2008), there has been comparatively little corpus research carried out on their undergraduate level writing, despite the high-stakes nature of this assessment. This paper explores a 170,000-word corpus of undergraduate assignments from first language (L1) Chinese students within 5 disciplines in UK universities, comparing this with a reference corpus of 580,000 words from L1 English students in the same disciplines (the majority of the data is extracted from the British Academic Written English corpus, Nesi, 2008). A keyword and key n-gram search is initially employed as a way in to uncovering differences in the writing of each student group; this is followed by searches for tagged non-linguistic items, and examination of concordance lines, collocates and dispersion plots to follow up the context of key items.

The paper focuses on two major differences between the student groups, namely the Chinese students' extensive use of visual elements such as tables, figures, images and diagrams, and their higher use of writing formatted as lists rather than as continuous prose. It is hypothesised that use of these features are strategies for L1 Chinese students who have to meet the challenge of producing multiple, extended pieces of writing in their second language. Presenting information using visual elements to support and extend ideas in the written language, and making points within a list format allow students to convey their thoughts clearly and effectively and in a more visually-oriented manner, while reducing the quantity of connected prose they have to produce.

Recent work within the field of multimodal analysis has highlighted the role of non-linguistic resources such as images and layout in all areas of communication (see Jewitt, 2009, for an overview). Exploring multimodality within corpus linguistics entails a focus beyond traditionally-privileged linear text, and this study explores the extent to which traditional corpus procedures can be used in the new 'trend' for multimodal corpora.

References

British Council. (2010). *China Market Introduction*. Retrieved 01/12/10, from http://www.britishcouncil.org/eumd-information-background-china.htm

Chuang, F.-Y., & Nesi, H. (2006). An analysis of formal errors in a corpus of L2 English produced by Chinese students. *Corpora, 1*(2), 251-271.

Mayor, B., Hewings, A., North, S., & Swann, J. (2007). A linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2. In L. Taylor & F. Falvey (Eds.), *IELTS Collected Papers: Research in Speaking and Writing Assessment. Studies in Language Testing* (Vol. 19, pp. 250-313): Cambridge University Press.

Hyland, K. (2008). Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41-62.

Jewitt, C. (2009). The Routledge Handbook of Mulimodal Analysis. London: Routledge.

Nesi, H. (2008). *BAWE: An introduction to a new resource.* Paper published in the Proceedings for The 8th Teaching and Language Corpora Conference, Location.

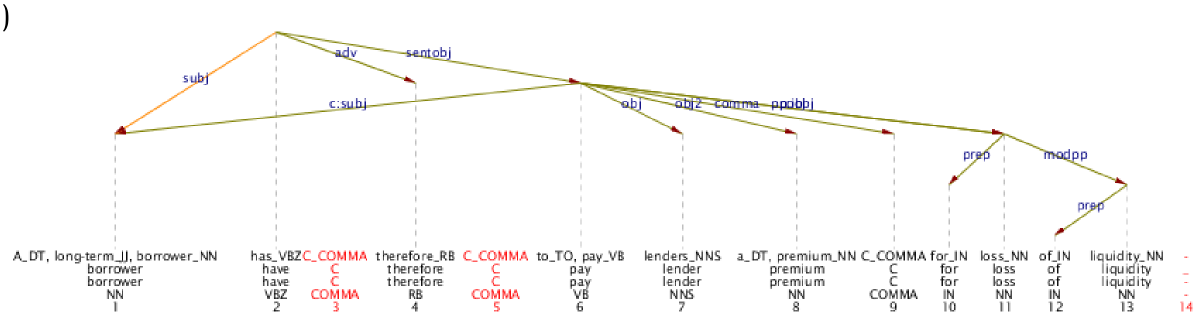
Note: The data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (formerly of the Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

BNC Dependency Bank 1.0

Hans Martin Lehmann, Gerold Schneider
University of Zurich

In this paper we present the first release version of a dependency bank for the British National Corpus (Aston & Burnard 1998). The BNC Dependency Bank was produced with Pro3Gres (Schneider 2008), a robust dependency parser with a handwritten grammar, which makes it ideal for experimentation in corpus linguistics. The parser produces automatic annotations, as in 1). For the whole BNC, the dependency bank contains over 87 million dependency relations, which are represented as individual arrows in 1).

1)


subj
adv
sentobj
c:subj
obj
obj2
comma
ppobj
prep
modpp
prep
A_DT, long-term_JJ, borrower_NN borrower borrower NN 1
has_VBZ have have VBZ 2
C_COMMA C C COMMA 3
therefore_RB therefore therefore RB 4
C_COMMA C C COMMA 5
to_TO, pay_VB pay pay VB 6
lenders_NNS lender lender NNS 7
a_DT, premium_NN premium premium NN 8
C_COMMA C C COMMA 9
for_IN for for IN 10
loss_NN loss loss NN 11
of_IN of of IN 12
liquidity_NN liquidity liquidity NN 13
- _ - - 14

We provide an overview of the annotation chain and discuss the problems and strategies involved. We also give a detailed description of the annotation scheme produced by Pro3Gres.

This project is based on fully automatic annotation, which – in our opinion – is the only viable approach for large corpora like the BNC. We present an evaluation of the parser performance. We argue that for many research questions the problems concerning parser performance can be contained. Indeed, many research questions cannot be tackled empirically without the amount of data accessible via automatic annotation. We exemplify and define the class of research questions that profit from a large-scale dependency bank.

We present strategies and methodology for extracting data for corpus-based and corpus-driven studies and consider the possibilities and limitations of the fully automatic annotation.

References

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

Aston, Guy & Burnard, Lou. 1998. *The BNC handbook exploring the British national corpus with SARA.* Edinburgh: Edinburgh University Press.

BNC Dependency Bank Online 1.0

Hans Martin Lehmann, Gerold Schneider
University of Zurich

In this software demonstration we present the core functionality of our interface to the BNC Dependency Bank, a syntactically annotated version of the British National Corpus. It is intended to provide corpus linguists with web-based access to the syntactically annotated BNC. The dependency annotation allows for direct access to the heads and dependents. Together with the graphical query tool this makes BNC Dependency Bank Online also suitable for teaching English linguistics.

We show syntactic queries with and without lexical constraints. We also present a tool for the analysis of lexical and syntactic types defined by underspecified queries, which permits a corpus-driven exploration of the syntax-lexis interface. We are confident to be able to present a publically accessible version of this project at the conference.

Like I said again and again and over and over. On the ADV1 and ADV1 construction in English

Magnus Levin, Linnaeus University
Hans Lindquist, Malmö University

Phrases and constructions have received increasing attention in recent years (e.g., Jackendoff 2008) and their relevance for various theories of language has been emphasized. For instance, Gries (2008:3–25) compares the role of "phraseologisms" in generative linguistics, cognitive linguistics, construction grammar and corpus linguistics, concluding that their importance can hardly be overestimated. In this paper we investigate an adverbial pattern which has so far been largely overlooked, namely ADV_1 *and* ADV_1, as in *again and again*, *by and by*, *over and over* and *through and through*. This construction has a number of uses, from the more or less transparent coordinated adverb phrases *again and again* and *over and over* to the negatively-charged fixed premodifier *out-and-out (lie)* and the fixed and non-transparent idiom *on the up-and-up* ('improving' (BrE) or 'honest' (AmE)). Adverbials are particularly interesting in phraseological studies, since, as Wray (2008:16) notes, many fixed phrases tend to occur in adverbial roles.

Diachronic data from the COCA and COHA corpora show that these patterns follow typical paths of change, such as a shift towards more abstract meanings, the fixing of collocates (as is typical for lexicalization (Brinton & Traugott 2005:105)), iconic variation (*again and again and again* referring to multiple repetitions) and grammaticalization (*by and by* > *baimbai* > *bai* as a Tok Pisin future marker being the prime example). Thus, for instance, *on and on* typically co-occurred with movement verbs (*walk/float*) in the 1800s, while in present-day English it mainly occurs either in the idiom *go on and on* usually expressing negative connotations about durative, non-physical activities, or in a verbless textual function ('and so on', as in *There have been workshops, task force reports, and on and on*). Similarly, *over and over* has been shifting from more literal meanings (*turning the letter over and over*) towards the more abstract meaning of (punctual) repetition, often with negative connotations (*saying over and over (again)*). Furthermore, *over and over* is being used increasingly without *again*, which suggests further lexicalization.

The results indicate that symmetrical ADV_1 *and* ADV_1 phrases in English over time have developed specialized meanings and show signs of metaphorization, lexicalization and grammaticalization. The analysis of this pattern thus provides insights into central processes of language change.

How could I write that? – revealing *of*

Kerstin Lindmark
Stockholm University

Prepositions are a notorious source of confusion both for learners of a new language and for native speakers learning to translate into their L1. In the case of English and Swedish, the fact that many prepositions are cognates further complicates matters. Cognate prepositions generally share several semantic features, but differ in their use, governed not only by the actual sense, but by conventions.

Especially treacherous is the pair "of" – "av". While "of", (representing 26% of preposition occurrences in the ESPC (English originals) and 20 or more different meanings (Garretson 2005), is used as a universal preposition, e.g., linking nouns forming one concept, Swedish "av" (11,3% of prepositions in the Swedish ESPC originals), cannot be used in this way. Neither is “av” used for expressing possession (Hammarberg & Koptjevskaja-Tamm 2003); instead, the genitive form of the possessor, another preposition, or a compound is used. However, the “N1 av N2”-construction does occur, although in other constructions.

Translating cognates causes special problems: (novice) translators may not be aware of the contrastiveness in the direction into their L1 and may get their implicit L1 competence (Paradis 2004, 2009) blurred by the L2 source construction. According to Shlesinger & Malkiel (2005) “the cognate is the first solution to be considered, and […] is rejected when the translator […] is convinced that the noncognate solution is superior”. The ability to make that kind of considerations is still lacking in inexperienced translators. Therefore, acquiring explicit L1 knowledge is necessary.

It is hypothesized that the “N1 av N2” construction is used in Swedish target texts even when this constitutes a violation of TL norms; and that such renditions will be more frequent the more inexperienced the translator is.

To confirm these hypotheses, the present study explores translation equivalents of “of” in beginners’ and professionals’ translations from English into Swedish, especially “N1 of N2”/“N1 av N2” constructions.

The material used is a corpus of translation students’ and patent attorney trainees’ translations, with source texts and model translations by professional translators. The English-Swedish Parallel Corpus (Altenberg & Aijmer 2002) serves as a reference corpus, complemented by the monolingual Swedish Stockholm-Umeå Corpus (Ejerhed & Källgren 1997) and the BNC Sampler (Burnard 1999).

Preliminary results indicate that the “N1 av N2” construction does occur in beginners’ translations in contexts where it is not used in Swedish originals, and more frequently in the patent attorney trainees’ texts than in university translation students’ texts.

References

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

Ejerhed, E., and Källgren, G. (1997.) *Stockholm Umeå Corpus version 1.0, SUC 1.0.* Umeå: Department of Linguistics, Umeå University.

Garretson, G., 2005: The meanings of English of: Uncovering semantic distinctions using a translation corpus, MA thesis, Boston University

Hammarberg, B. & M. Koptjevskaja-Tamm, 2003: Adnominal possession: combining typological and second language perspectives. In Giacalone Ramat, A. (ed.), *Typology and Second Language Acquisition.* Berlin: Mouton de Gruyter, 125-180.

Paradis, Michel, 2004: A Neurolinguistic Theory of Bilingualism. Benjamins

Paradis, Michel 2009: Procedural and Declarative Determinants of Second Languages. Benjamins

Shlesinger, M. & B. Malkiel. 2005. Comparing Modalities: Cognates as a Case in Point. *Across Languages and Cultures* 6:2. 173-193.

Electronic resources:

Burnard, L. 1999: Users Reference Guide for the BNC Sampler, http://www.natcorp.ox.ac.uk/corpus/sampler/

Website of ESPC (English-Swedish Parallel Corpus). http://www.englund.lu.se/content/view/66/127/. Visited 14/5/2009.

Språkbanken website. http://spraakbanken.gu.se. Visited 15/7/2009.

Generally or *In most cases*? Syntactic realization of multi-word stance markers

Anne Li-E Liu
University of Nottingham

This paper investigates how native speakers of English (NS) and L2 writers show their stance/attitude by employing multi-word stance markers in their writing. Stance markers are single words like adverbs (undoubtedly, definitely) or verbs (think, presume); yet, the same notion 'stance marker' can be represented via various syntactic realization. For example, in expressing a writer or a speaker's stance with the single-word marker *frankly*, other alternatives are to employ multi-word versions of this concept. This includes the prepositional phrase--*in all frankness*, the *-ing* clause--*frankly speaking*, and the finite clause--*if I may be frank*. Biber et al. (1999) look at the syntactic forms of stance markers (their term: stance adverbials) in four registers and conclude that single-word adverbs occur relatively frequently; yet they also state that "prepositional phrases are the second most common form in news and academic prose" (1999: 862). Following that, literature that examines the stance markers mainly focuses on the distribution of various forms (Fang 2006; Zhen 2008). Note that other than single-word markers, the rest of the syntactic categories can be clustered and termed as multi-word stance markers. In showing the concise manner of the language used, other than employing the single-word marker *briefly*, a writer has various multi-word stance markers like *in short*, *to put it briefly*, *in a word*, and *simply put* at his disposal. A question that is worth pursuing is to what extent would a writer favour one over the other? This study aims at exploring this aspect and providing a more fine-grained description with regard to how NS and L2 writers use multi-word stance markers in their writing.

Twenty-one multi-word stance markers in four groups are included, markers that show *unexpectedness*, *manner*, *respect*, and *generalness*. To provide a direct comparison, I search these markers in International Corpus of Learner English (ICLE) and the spoken section and the written prose of BNC, BNC-S and BNC-W, respectively. Both qualitative (manual examination) and quantitative methods (raw frequency, normalized frequency and the log-likelihood score) are adopted in data collection and analysis. A different pattern of multi-word stance markers is found between L2 writers and the NS. For example, 42% of the markers used by L2 writers in showing *unexpectedness* are the non-finite clause realization, *to my//his/her/our/their surprise*. Such tendency is also found in the BNC-S. NS writers, on the other hand, prefer the single-word marker, *surprisingly* and multi-word markers are sparsely seen.

References

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

Fang, A.C. (2006). A Corpus-Based Empirical Account of Adverbial Clauses across Speech and Writing in Contemporary British English. In LNCS 4139/2006: *Advances in Natural Language Processing*. Berlin and Heidelberg: Springer. pp 32-43.

Zhen, C. (2008). Corpus-based Study of the Stance Adverbials in Chinese-English Political Texts, *UCCTS 2008*, China, Hangzhou.

Male and female swearing in the contextually governed texts in the BNC

Magnus Ljung
University of Stockholm

The present paper is a sequel to my paper *Men, women and swearing in the BNC* which was read at the panel on English-language swearing held at SS18 (the Sociolinguistics Symposium 18) held at the University of Southampton September 1- 4, 2010. The Southampton paper compared male and female swearing in the demographically sampled texts and the contextually governed texts in the spoken component of the BNC. The results of the comparison suggested that, unlike the demographically sampled texts, the contextually governed texts contained less female than male swearing. This finding was unexpected and the present paper is an attempt to test the validity of this result and to explore it further by studying the effect on the use of swearing among males and females of variables such as the age and social class of the speakers and the different types of discourse involved in the contextually governed texts.

Reference

Ljung, M. (2010). *Men, women and swearing in the BNC*. Paper read at SS18.

Looking into the history of *namely*: A story of ruthless competition

María José López-Couso
University of Santiago de Compostela

A look at the early history of *namely* reveals that this form arose in the late twelfth century as a particulariser (cf. Quirk et al. 1985: 604), with the meaning 'particularly, especially, above all' (OED s.v. *namely* adv. 1; MED s.v. *nam(e)li* adv. 1), and that it was only in Late Middle English that it developed its present-day function as an optional marker of expository apposition (cf. Quirk et al. 1985: 1307ff; Meyer 1992: 97), meaning 'to wit, that is to say, videlicet' (OED s.v. *namely* adv. 3; MED s.v. *name(e)li* adv. 2). Early instances of *namely* in these two functions are given in (1) and (2).

(1) (1)*Sunnedei ah efri cristenne Mon* ***nomeliche*** *to chirche cume* (c1175, *Lambeth Homilies* 139).
(2) (2)*In that the feende repaireth moste, bothe in man and woman,* ***namely,*** *when they be in grete ire* (c1450, *Merlin* 8).

In its original particularising function, *namely* faced stiff competition from a host of adverbs and phrases which developed in the course of Middle and Early Modern English with a similar meaning and use, such as *(e)specially*, *in (e)special*, *particularly*, *in particular*, *principally*, *chiefly*, and *notably*, among others. After a period of variation, these new formations, mostly derived from Romance bases, ousted *namely* from the inventory of English particularisers by the end of the Early Modern English period. On the other hand, once *namely* developed its derived meaning and function as a marker of expository apposition, it came to compete in usage against a number of rivals in its new domain, among them *to wit*, *that is to wit(ting)*, *to understand*, *id est*, *scilicet*, *videlicet*, *that is*, and *(that is) to/for to/at say* (cf. Nevanlinna & Pahta 1997; Pahta & Nevanlinna 1997, 2001).

My aim in this paper is to explore these two instances of linguistic competition with opposite results in which *namely* has been involved in its lifetime, looking into the factors which may have contributed to its failure and its success in these two cases of competition for the same semantic space. Special attention is paid to (i) the potential influence of cognitive motivations (e.g. ease of processing, degree of morphological and semantic transparency), (ii) the textual distribution of the variants, and (iii) the syntactic correlates of the semantic change from particulariser to optional appositive marker. Evidence is drawn from a variety of sources, including the *Helsinki Corpus of English Texts* and ARCHER 3.1.

References

ARCHER 3.1 = *A Representative Corpus of Historical English Registers*. 2006. Northern Arizona University, University of Southern California, University of Freiburg, University of Helsinki, and Uppsala University.

HC = *The Helsinki Corpus of English Texts*. 1991. Helsinki: Department of English.

MED = *Middle English Dictionary*, ed. Hans Kurath, Sherman M. Kuhn & Robert E. Lewis. Ann Arbor: University of Michigan Press.

Meyer, Charles F. 1992. *Apposition in Contemporary English*. Cambridge: C.U.P.

Nevanlinna, Saara & Päivi Pahta. 1997. "Middle English non-restrictive apposition with an explicit marker." In Jacek Fisiak (ed.). *Studies in Middle English Linguistics*. Berlin & New York: Mouton de Gruyter: 373-401.

OED = *The Oxford English Dictionary on CD-ROM*, ed. John A. Simpson & Edmund S.C. Weiner. 2nd edn. Oxford: O.U.P.

Pahta, Päivi & Saara Nevanlinna. 1997. "Re-phrasing in Early English: The use of expository apposition with an explicit marker from 1350 to 1710." In Matti Rissanen, Merja Kytö & K. Heikkonen (eds.). *English in Transition. Corpus-based Studies in Linguistic Variation and Genre Styles*. Berlin & New York: Mouton de Gruyter: 121-183.

Pahta, Päivi & Saara Nevanlinna. 2001. "On markers of expository apposition." *NOWELE* 39: 3-51.

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

From reduction to conventionalization: *gonna* versus *going to*

David Lorenz
Universität Freiburg

In studies and models of grammaticalization, the focus has often been on how a lexical item gradually acquires new meanings and (grammatical) functions (e.g. Heine 2002, Diewald 2002). It is also well known that grammaticalization comes with a rise in frequency and, potentially, phonetic reduction (Bybee 2006). But the story does not end here. A phonetically reduced form of the grammaticalized item may become conventionalized and thus become a competing variant of its source form.

I argue that this is the case with the future marker *gonna* in English.

Gonna is of course a reduced variant of the semi-modal *going to.* It is also conventionalized in that it is not restricted to rapid speech and has a standard ("correct") spelling.

My study is based on data from the Santa Barbara Corpus of Spoken American English (SBC) and the Corpus of Contemporary American English (COCA). It shows that diachronically *gonna* is winning out against *going to* (in apparent time), at least in spoken language. An analysis of contexts and phonological realization shows that contexts expected to favor reduction (e.g. pronominal subjects) do not account for the use of *gonna*, while other phonetic reduction (e.g. [aɪmə], "I'm gonna") is tied to these contexts. This suggests that the changing variation between *gonna* and *going to* cannot be attributed to a tendency to phonetic reduction. Rather, speakers make a choice between two different constructions.

In the context of grammaticalization this study sheds light on the transition from automization and reduction (pertaining to the individual speaker) to the conventionalization of a new variant (which pertains to the language community as a whole). It also yields new insights into the ongoing restructuring of the English modal system.

References

Bybee, Joan. 2006. "From usage to grammar: The mind's response to repetition". *Language,* Vol. 82, No. 4. 711-733.

Davies, Mark. 2008-. *The Corpus of Contemporary American English (COCA)*: 410+ million words, 1990-present. Available online at http://www.americancorpus.org.

Diewald, Gabriele. 2002. "A model for relevant types of context in grammaticalization." *New Reflections on Grammaticalization*, ed. by I. Wischer and G. Diewald. Amsterdam, Philadelphia: John Benjamins. 103-120.

Du Bois, John W. and Robert Englebretson. 2003. *Santa Barbara Corpus of Spoken American English.* University of California, Santa Barbara Center for the Study of Discourse.

Heine, Bernd. 2002. "On the role of context in grammaticalization." *New Reflections on Grammaticalization*, ed. by I. Wischer and G. Diewald. Amsterdam, Philadelphia: John Benjamins. 83- 101.

Clause-final 'man' in Northeast English

Kathrin Luckmann
University of Duisburg-Essen

Clause-final *man* has been rather unambiguously attributed to the speech of 'Geordies' or North-Easterners in general. But there has been little research of the feature.

This paper investigates the following questions:

i) the status of *man* as a form of address (Beal 2008a: 400; British Library), or pragmatic marker;
ii) its function according to the prior categorization as expressing 'annoyance or impatience' (Beal 2008a: 400; examples 1 and 2 illustrate this function), its use to 'attract attention or establish solidarity' (British Library; example 3 illustrates this function), or the macro function of marking regional identity;
iii) and the question of prominence of clause-final elements.

These possibilities concerning clause-final *man* are tested across evidence from the language used in several episodes of the popular ITV series *Auf Wiedersehen, Pet* from 1983, in which the three bricklayers, Dennis, Neville and Oz, leave Newcastle to work on a building site in Düsseldorf.

(1) I divn't gan for holidays *man.* I wish I could. (Beal 2008a: 400; my emphasis)
(2) You don't, like, go asking them for favours, *man.* (*Auf Wiedersehen, Pet*: 15'53)

(3) A: It’s too late for that now, Den. I just rang the missus and told her I’m coming.
B: Well, was it her letter, like?
A: No it wasn’t her. She’s put no pressure on at all.
B: It’s me *man* Dennis, I hate it here, I admit. (*Auf Wiedersehen, Pet*: 41’35)

My study indicates that *man* in many instances has a more interpersonal quality and can, in these instances, more fittingly be described as pragmatic marker than form of address (see example 3 above).

In addition, the paper shows how *man* is employed to mark regional identity by the speakers. It seems to be a distinguishing feature which serves to mark out Geordie speakers as just that, dissociating the Geordie ‘brickies’ linguistically from their Liverpudlian, West Midland and Cockney colleagues in *Auf Wiedersehen, Pet*.

The clause-final position, in this variety, is prosodically prominent as well as salient because of a pattern that sets it apart from other varieties, including other Northern varieties. The intonational pattern involved is the so-called *Urban Northern British Rise* with a ‘rise-plateau intonation in declarative sentences’ (Beal 2008b: 140-1; see Kerswill 2002: 103; who argues for non-prominence in this position).

“Who’s afraid of ...... what?” – in English and Portuguese

Belinda Maia, Universidade do Porto
Diana Santos, University of Oslo

Fear is generally accepted as a primary emotion in studies on the relationship between cognition and emotion. In this paper we shall use monolingual and parallel corpora to explore ways in which the English and Portuguese languages reflect descriptions of this emotion and its relation to cognitive processes.

The emotion lexicon uses verbs like *to fear* where, to use systemic-functional grammar terminology, the Subject is Senser and the Object is Phenomenon. However, in English, the use of these verbs is actually less frequent than structures with the Senser + *be* + past participle or adjective (*afraid /frightened/ anxious*) + a preposition like *of* or *about* + the Phenomenon (rather than the *by* phrase associated with a passive), or by a *that* clause. These constructions express varying degrees of conscious analysis of the Phenomenon causing fear. In Portuguese, the main means of expressing fear is through the verbs *recear* and *temer,* the expressions *ter medo*, and *estar com medo*, literally *to have fear*, and *to be with fear,* and verbs like *assustar* (*frighten*), that behave like *frightened*, but can also function in a reflexive type manner (*assustar-se*), see Maia (1996).

The use of this lexicon is interesting in that the meaning of the lexical words will depend a lot on the Phenomenon involved. Compare, for example, *afraid of heights* with *afraid of white space*. Then there is the difference between *afraid of* and *afraid that* and the possibilities in Portuguese of *ter medo de, recear,* and *temer* as well as the different distribution in the two languages of nominal vs. verbal descriptions of fear.

We shall first show how the lexicon of fear behaves in monolingual corpora both for English and for Portuguese. We shall then use parallel corpora to see how these expressions are translated and how their translation casts light into interesting differences between the two languages.

Research so far suggests that the FEAR lexicon and its associated syntax offer interesting insights into how human beings verbalize their experience of fear and use it to express other communicative functions. However, there are nuances of meaning that are only noticeable when subjected to translation, as Stig Johansson showed with *loving* and *hating* in English and Norwegian.

Reference

Maia, Belinda. "A Contribution to the Study of the language of Emotion in English and Portuguese". PhD thesis, Porto: FLUP, 1996.

VOICE XML: Xtending VOICE

Stefan Majewski
University of Vienna

The Vienna-Oxford International Corpus of English (VOICE) is a corpus of English as a lingua franca (cf. Seidlhofer 2005). It features transcripts of highly interactive spoken communication and provides a rich set of meta-data. Since mid 2009, VOICE has been freely available to a continuously growing user base via the web application VOICE Online (cf. VOICE Project 2009). While the web application focuses on patterned lexical search, the release of the entire XML resource for VOICE allows for more flexible qualitative and quantitative research using all available annotation and meta-data in VOICE.

The XML edition of VOICE is a corpus encoded in a data-format based on the recommendations of the Text Encoding Initiative (TEI) (cf. Burnard and Bauman (eds.) 2008). The format is defined as a customisation of the TEI P5 Guidelines and consists of a selection of elements from relevant modules of the TEI. This definition is specified and documented using TEI's own schema specification language (ODD). From this central definition, all required schemas for the formal validation of VOICE XML are derived. The VOICE Project decided on a data representation in a hierarchical document format. Therefore, as the mark-up is in line with the annotated text, the annotation is immediately accessible from the transcribed material.

The proposed paper presents the structure and research possibilities with VOICE XML. It takes the requirements of annotation, transcription and research as its conceptual starting point. From there, the main structure of the corpus is developed. This includes the representation of stratification as well as the available annotation layers for linguistic and paralinguistic features and the available meta-data on corpus, event and speaker level. Special emphasis is placed on the representation of temporal aspects (e.g. overlapping speech) within the highly interactive data comprised in VOICE. Furthermore, access techniques are proposed to exemplify research possibilities based on VOICE XML and similar TEI-based corpora. The proposed methods are suggested as guidance to finetune the researcher's choice of methods in accordance with the design criteria of the corpus. Additionally, the paper discusses generic mechanisms to add custom layers of annotation to the corpus as a resource and thus to individually extend the corpus resource. These mechanisms are currently used in the VOICE Project to add syntactic annotation (cf. Osimk (forthc.)), but are expected to be generically applicable to capture and reuse individual research findings as additional layers of corpus annotation.

References

Burnard, Lou; Bauman, Syd (eds.). 2008. *TEI P5: Guidelines for Electronic TextEncoding and Interchange.*

Osimk, Ruth. (forthc.). "Evaluating the applicability of existing POS practices for acorpus of English as a lingua franca (VOICE)". In Hoffmann, Sebastian;

Rayson, Paul;Leech, Geoffrey (eds.). *Corpus linguistics and variation in English: Focus on Non-native Englishes (Proceedings of ICAME 30)*. Helsinki: Research Unit for Variation,Contacts and Change in English (VARIENG), University of Helsinki.

Seidlhofer, Barbara. 2005. "English as Lingua Franca". *ELT Journal.* 59, 339-341.

VOICE Project. 2009. *VOICE – Availability.* http://www.univie.ac.at/voice/page/corpus_availability (2010-11-27).

The "have" construction

Michaela Martinková
Palacký University

English constructions with a raised object after *have* followed by the infinitive or the *ing* form are by Quirk et al. (1985:1205,1206) ranked among the complex transitive kind of complementation with coercive meaning, and further considered in comparison with the corresponding 'existential' *there* constructions. Unlike these, the subject in the *have* construction has "considerable involvement in the existential proposition", though "we cannot specify what that involvement will be" (Quirk et al. 1985:1411). The subject of *have* can be involved as either the experiencer or initiator. Stating formal criteria to differentiate between the experiential reading (*John had a funny thing happen to him*) and the causative reading (*John had a man cut the grass*) is problematic (Austin 2004:77). Macháček's understanding of *have* as an item expressing "a relation of inclusion in the subject" (1965:13), and thus potentially creating "the sphere of the subject's concern" (1965:27) allows one to see causation as a special kind of this "concern".

This paper focuses on the *have*+object+V*ing* construction and compares it with Czech translations (the InterCorp Project). It turns out that if the corresponding variant with *there* exists (the object is an NP with indefinite reference functioning as the focus of the sentence), the subject's concern is often lost in the Czech translation. This happens in sentences with transitive verbs or verbs with prepositional objects, where the subject of the whole sentence, referring to the one with concern, is co-referential with the object of this verb, and thus obligatory in Czech (*But **he** had an excellent team working on **him**. Ale operoval **ho** skvělý tým* "but – operated – him – great team"). In sentences with intransitive verbs, or those where the object of *have* is an NP with definite reference or a personal pronoun, the participant with the concern can be expressed if a Czech verb with a dative construction is introduced (*So useful to have him swooping around like an overgrown bat. Vždycky **mi** bylo velice vhod, když* "always – to me – was –very – suitable – when"). The Czech 'non-attached' dative, often mentioned in linguistic literature (Poldauf 1964:250) as a means of introducing the participant with concern (***I** won't have you coming here Nechoď **mi** sem* "don't – go – to me – here") is unduly under-represented in the Czech translations. The causative meaning, which is only an inference from the context, tends to be explicitly expressed in Czech through causative constructions, prefixes, or prepositional phrases.

Combining corpus and experimental data in second language acquisition research

Amaya Mendikoetxea, Universidad Autónoma de Madrid
Cristóbal Lozano, Universidad de Granada

This paper shows how corpus data and experimental data can be combined to gain an insight into the processes that shape and develop L2 acquisition. Using the International Corpus of Learner English (ICLE, Granger et al. 2002) we briefly report on a corpus study on subject position in L1 Spanish – L2 English, which revealed that subject position in L1 and L2 English is constrained by the same principles (see Lozano and Mendikoetxea 2010). These principles have to do with verb type (unaccusative/unergative) as well as with the status of the subject (information status and weight). This study shows that large and well constructed corpora are powerful tools for our understanding the processes that constrain L2 production.

The results of the corpus-based study revealed that learners had difficulties in identifying what element should occur preverbally in structures with postverbal subjects. As a follow-up, an online experiment was designed to test learners' knowledge of the types of elements which may appear preverbally in those contexts. Learners had to judge, on a five-point Likert scale, the acceptability of 32 contextualised sentences with postverbal subjects containing 8 different verbs. Crucially, these sentences were structurally similar to those extracted from the corpus: 4 of those verbs were our top inversion verbs in the corpus *(exist, appear, begin and come),* while the other four where verbs for which no inversion structures were found *(talk, work, play and speak).* Variables concerning the status of the subject (focus and heavy) were controlled for. The preverbal position matched what we had found in the corpus: sentences containing *it, there, PP* or a *zero* subject. In total, over 250 L1 Spanish-L2 English learners, at all levels of proficiency, participated in the experiment.

Given the design of the experiment and the high number of learners, the results show a very robust pattern, which mostly matches the one obtained in the corpus study. Thus, we can talk about converging evidence, but there are also some interesting (and consistent) deviations which are commented on in detail. In the conclusions we argue that using converging evidence to triangulate results is paramount in current second language research: corpus-based results can, and should, be validated against corpus-external findings, and combining naturalistic and experimental data is crucial to gain insight into the relation between the two types of data (see Gilquin & Gries 2009).

References

Gilquin, Gaëtanelle & Gries, Stephan (2009). Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1). 1-26.

Granger, S., Dagneaux, E., & Meunier, F. (Eds.). (2002). *International Corpus of Learner English* (inc. CD ver 1.1). Louvain: UCL Presses Universitaires de Louvain.

Lozano, C. & Mendikoetxea, A. (2010). Interface conditions on postverbal subjects: a corpus study of L2 English. *Bilingualism: Language and Cognition*, 13(4):475-497.

The Longitudinal Database of Learner English LONGDALE: focus on tense and aspect errors

Fanny Meunier, Damien Littré
Centre for English Corpus Linguistics UCL

Our study aims to illustrate how the analysis of a longitudinal learner corpus can help shed light on the cognitive processes at play in second language acquisition and provide useful information for pedagogical applications (Meunier 2008). We examine the prevalence and trajectory of tense-aspect errors in essays written by French-speaking learners of English – from intermediate to advanced levels – over a period of three years.

We analyze the written productions of 38 students belonging to the first cohort of subjects included in the *Longitudinal Database of Learner English (LONGDALE*[1]*).* After a brief presentation of LONGDALE, we present an overview of the general patterns characterizing the evolution of tense-aspect errors in L2 writing over a period of three years, we investigate whether some categories of tense-aspect errors still pose difficulty to advanced learners. We compare categories of tense-aspect errors (e.g. confusion between the present perfect and past simple, present continuous and present simple, etc.) to give a comprehensive account of the tense-aspect errors found in L2 writing and we study the developmental patterns that lead to this outcome. We particularly focus on whether time and proficiency have an influence on the number and distribution of errors.

We also discuss some pedagogical implications of our findings. The role of cognitive linguistics in SLA and the potential of using a cognitive approach to teaching English grammar in ESL/EFL settings has convincingly been put forward a.o. by Robinson and Ellis (2008) and Radden & Dirven (2007); whilst we concur with the authors, we will argue that taking into account the results of learner corpus research makes it possible to better target the pedagogical treatment and meet learners' attested needs.

References

Meunier, F. (2008) *Corpora, cognition and pedagogical grammars : An account of convergences and divergences* . In De Knop, S. and T. De Rycker (eds) Cognitive Approaches to Pedagogical Grammar. Berlin: Mouton de Gruyter, 91-119.

Radden, G. and Dirven, R. (2007) *Cognitive English Grammar.* Amsterdam: John Benjamins.

Robinson, P. and Ellis, N. (2008) *Handbook of Cognitive Linguistics and Second Language Acquisition.* New York/London: Routledge.

[1] For more information on the LONGDALE project, see http://www.uclouvain.be/en-cecl-longdale.html

Politeness conventions in Old English and Old Norse

Charles Meyer, University of Massachusetts Boston
Viola Miglio, University of California Santa Barbara

In his discussion of Old Norse in Danelaw England, Townend (2002: 182) reaches the conclusion that 'the available evidence points fairly unequivocally to a situation of mutual intelligibility of speakers of Norse and English in the Viking age'. However, while structurally the languages may have been similar, pragmatically the situation is more complicated, especially if politeness conventions for both languages are compared. Both languages employed insults in flyting-type of exchanges. However, their use of address terms are really quite different. In fact, the address terms discussed in Kohnen (2008) do not typically occur in Old Norse, a more secular society, where the Christian conventions found in OE do not appear, being substituted by the simple use of first names. In our presentation, we will use Leech's (1983) notions of tact, approbation, generosity, and modesty (all part of his politeness principle) to compare conventions of politeness in Old English and Old Norse.

We specifically analyse stylized sexually-based insults or suggestive double-entendres, as well as ritualized turns of phrases encountered in heroic poetry such as *Beowulf* or in the OE riddles and compare them to ON terminology found in the Old Icelandic family sagas and in flyting-type of poetry such as *Lokasenna*. We track the terminology and turns of phrases used throughout the available corpora (such as the DOE for Old English, the IcePaHC for Icelandic, as well as the texts available for online search at northvegr.org, and sagadb.org).

Kohnen (2008:155) concludes that 'politeness as face work may not have played a major role in Anglo-Saxon England' or that, at least, it may have been expressed through different linguistic means than those he researched. We notice that this is not the case in the context of the flyting – where ritualized verbal sparring is expected and therefore acceptable. In the Old Icelandic sagas we also find the use of politeness principles: women, for instance, use highly ritualized exchanges where insults are not openly expressed, but only hinted at as part of directives to spur someone to action (revenge), or to express disagreement and disappointment facing a *fait accompli* (arranged marriage). These exchanges express a type of required face work – indispensible for women to goad or chide men.

We conclude that politeness considerations are relevant also for medieval societies even if their implementation needs to be analysed in the context of the medieval cultural world.

References

Leech, G. (1983) *Principles of Pragmatics*. London: Longman.

Kohnen, T. (2008) "Linguistic politeness in Anglo-Saxon England? A study of Old English address terms." *Journal of Historical Pragmatics* 9.1: 140-158.

Townend, M. (2002) *Language and History in Viking Age England*. Turnhout, Belgium: Brepols.

Discourse communities and their writing-styles: A case study of Robert Boyle

Lilo Moessner
RWTH University Aachen

In the present study it is assumed that individuals can be members of more than one discourse community, and that a shared object of study need not form a discourse community (Swales 1990: 24-32). Disciplines are understood as social spaces in which individuals or groups interact according to certain conventions and with the aim of advancing and spreading knowledge. This qualifies them as discourse communities. The producers of disciplinary discourse must meet the expectations of their disciplines, and it can be expected that expert members of a discipline will do so with more ease than novices in the field. Yet it is also the experts of a discipline whose power position allows them to transcend disciplinary boundaries and use individual writing-styles, which may even become the starting-point of new disciplinary conventions (Hyland 2009).

Within this theoretical frame-work I propose to analyse the writing-style of Robert Boyle, who is best known as a natural philosopher and one of the founders of the Royal Society. He was an expert member of the natural sciences, to which he contributed in the form of letters, of articles in the *Philosophical Transactions*, and of monographs. His contributions to the field of medicine are often overlooked in evaluations of Boyle's work (Hunter 1997: 322). The question to be tackled is whether the disciplinary pressures were big enough for Boyle to adopt two distinct writing-styles, or whether his authority in one or both disciplines allowed him to deviate from disciplinary conventions and create an authorial identity.

My study will be based on two parallel corpora of about 25,000 words each, compiled from Boyle's scientific and medical treatises. They will be analysed with the method of multidimensional analysis. The results of both sub-corpora will be compared to a scientific and a medical control corpus of about the same size. If Boyle's values are in line with those of other members of the same discipline, Boyle's writing-style follows the disciplinary conventions. If Boyle's values are different from those of other members of the same discipline, this can theoretically be an indication that he was not fully acquainted with the discourse conventions of the disciplines or that he creatively deviated from them. This is the result which my earlier research (Moessner 2009, Moessner submitted) makes me expect. If Boyle's writing-style changed the discourse conventions of the disciplines involved will be checked by comparison with control corpora from the middle of the 18th century.

References

Hunter, Michael. 1997. "Boyle versus the Galenists: a Suppressed Critique of Seventeenth-Century Medical Practice and its Significance". *Medical History* 41:322-361.

Hyland, Ken. 2009. Constraint vs Creativity: Identity and Disciplinarity in Academic Writing". *Commonality and Individuality in Academic Discourse*, Maurizio Gotti (ed.), 25-52. Bern, etc.: Peter Lang.

Moessner, Lilo. 2009. "The influence of the Royal Society on 17th-century scientific writing". *ICAME Journal. Computers in English Linguistics*. No 33: 65-87.

Moessner, Lilo. Submitted. The Rise of Disciplinary Identity.

Swales, John. 1990. *Genre Analysis. English in academic and research settings*. Cambridge: Cambridge University Press.

Text classification of the BNC using corpus and statistical methods

Ghada Mohamed
Lancaster University

This presentation demonstrates a new statistical methodology for establishing categories within a text typology. Although there exist many different approaches to the classification of text into categories, my study will fill a gap in the literature as most work on text classification is based on features external to the text such as the text's purpose, the text producer's intentions, and the medium of communication (see, for instance, Reiss 1976; Welirch 1976). Text categories that have been set up based on some external features are not linguistically defined (Biber 1989: 5). In consequence, texts which belong to the same type are not necessarily similar in their linguistic forms. Even Biber's (1988) linguistically-oriented work was based on externally-defined registers.

In this presentation I will show how a text typology, based on similarities in linguistic forms, can be developed using a multivariate statistical technique, namely cluster analysis. In this study, this technique was implemented using R statistical package. There are two reasons for using statistical software: (a) the large number of texts to be classified in the BNC (British National Corpus) and (b) the exhaustive list of linguistic features used as underlying variables for classifying the texts. The linguistic features used include personal pronouns, passive constructions, prepositional phrases, nominalization, modal auxiliaries, adverbs, and adjectives.

Computing a cluster analysis based on this data is a complex process with many steps. At each step, several alternative techniques are available. Choosing among the available techniques is a non-trivial decision, as multiple alternatives are in common use by statisticians. I will demonstrate how a process of trial and error was used to test several combinations of clustering methods, in order to determine the most useful/stable clustering combination(s) for use in the classification of texts by their linguistic features. The stable results obtained from this trial and error process were then validated using three validation techniques available in cluster analysis, namely the cophenetic coefficient, the adjusted Rand index, and the AU *p*-value.

Cluster analysis, if used with caution, is a powerful tool for structuring the data. The way it has been implemented in this study constitutes an advance in the field of text typology.

References

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

Biber, D. (1989) A typology of English texts. *Linguistics, 27*, 3-43.

Reiss, K. (1976) *Texttyp und Übersetzungsmethode. Der operative Text.* Kronberg: Scriptor.

Werlich, E. (1976) *A Text Grammar of English.* Heidelberg: Quelle and Meyer.

Order and law? Degrees of reversibility in English binomials

Sandra Mollin
University of Heidelberg

Binomial expressions have been defined by Malkiel (1959) as two words of the same form class, linked by a conjunction. Examples he gives are *odds and ends*, and *cold and snow*. While there has been no shortage of research on binomials in the past, this has almost exclusively focused on those binomials that are irreversible (a notable exception is Gustafsson 1976), and on the ordering principles lying behind the specific order that became fixed in these expressions (e.g. Cooper/Ross 1975, Fenk-Oczlon 1989, Benor/Levy 2006).

However, as this paper will show, only a minority of binomials is irreversible. Among the binomials occurring frequently (more than fifty times) in the BNC, only 18% display a fixed order. Binomials are distributed on a cline from irreversible to entirely reversible, with the majority of items exhibiting a preference for a certain order without excluding occurrences in the reverse order. As an example, *law and order* is irreversible in the BNC, whereas *values and beliefs* respectively *beliefs and values* occur in equal numbers, making this an entirely reversible binomial. *Time and money* is an example of a reversible binomial with a strong preference for one order, with 90% of cases occurring in just this order and only 10% of cases occurring as *money and time*.

Thus, this paper presents the first corpus-based analysis of binomials regarding degrees of reversibility, introducing the calculation of a reversibility score. A taxonomy of binomials is provided categorizing binomials according to both degree of reversibility and idiomaticity (idiomatic binomials are found towards the more 'fixed' end of the reversibility cline). Furthermore, the distribution of word classes will be discussed, since pairs of adjectives and adverbs are more likely to appear at the reversible end of the cline, whereas pairs of nouns, verbs, conjunctions and prepositions are more frequent among the irreversible binomials.

Finally, the paper will conclude with an outlook on how aware learners are of the differing degrees of reversibility of binomials. Over 100 advanced learners of English at a German university participated in an ordering judgment task. The results reveal that the learners were often not aware of subtle degrees of reversibility, but that they were aware of the order in irreversible binomials and binomials with a strong order preference.

References

Benor, Sarah Bunin & Robert Levy. 2006. "The chicken or the egg? A probabilistic analysis of English binomials." In: *Language* 82: 2, 233-277.

Cooper, William E. & John Robert Ross. 1975. "World Order." In: Robin E. Grossman, L. James San & Timothy J. Vance (ed.). *Papers from the Parasession on Functionalism*. Chicago: Chicago Linguistic Society, 63-111.

Fenk-Oczlon, Gertraud. 1989. "Word frequency and word order in freezes." In: *Linguistics* 27: 1, 517-556.

Gustafsson, Marita. 1976. "The frequency and "frozenness" of some English binomials." In: *Neuphilologische Mitteilungen* 77, 623-637.

Malkiel, Yakov. 1959. "Studies in irreversible binomials." In: *Lingua* 8, 113-160.

Periphrastic *do* in eighteenth-century correspondence: a sociolinguistic study

Arja Nurmi
University of Helsinki

This study follows the final stages of regulation in the use of periphrastic DO. During the eighteenth century, DO was already established in the patterns familiar from Present-day English: it was used in negation, inversion, code and emphasis. There was still variation in its use in negation and inversion, however, and it continued to be used in affirmative statements as a semantically empty tense carrier. The purpose of this paper is to trace the last stages of the s-curve in the spread of DO in negative statements and the decline of DO in affirmative statements. There are not enough interrogatives for a quantitative study.

The material used is the Corpus of Early English Correspondence Extension (CEECE), 2.2 million words of personal correspondence from 1681-1800. The study continues from where Nurmi (1999) left off, and complements the picture established in Tieken-Boon van Ostade (1987), which looked at the register variation of DO in a larger variety of texts but a notably smaller corpus. The main focus of the study is the sociolinguistic variation evident in the use or non-use of DO, but certain linguistic constraints are also tracked, most notably the *know*-group of verbs identified by Ellegård (1953) as particularly averse to adapting DO in negative statements.

References

Ellegård, Alvar (1953). The Auxiliary do. The Establishment and Regulation of Its Use in English. Stockholm: Almqvist & Wiksell.

Nurmi, Arja (1999). A Social History of Periphrastic do. Helsinki: Société Néophilologique.

Tieken-Boon van Ostade, Ingrid (1987). The Auxiliary do in Eighteenth-century English: A Sociohistorical-linguistic Approach. Dordrecht: Foris.

Contrasting writer's epistemic stance in English and Spanish opinion columns: a corpus-based study

María Pérez-Blanco
University of León

The present paper deals with the expression of writer's epistemic stance in English and Spanish opinion columns from a contrastive viewpoint. Opinion columns are generally characterized by the expression of the author's positioning or commitment towards the truth of the information presented in the text. Epistemic stance markers can mark certainty or doubt, actuality, precision or limitation; they can also comment on the source of the information (Biber et al. 1999). The present study will focus on the linguistic markers of certainty and doubt in a broad sense as in Biber and Finegan (1989). The categories analyzed will be: (1) certainty adverbials, (2) certainty adjectives, (3) certainty verbs, (4) doubt adverbials, (5) doubt adjectives and (6) doubt verbs. The principal focus of this paper is to explore the quantitative and qualitative similarities and differences in the use of epistemic markers between English and Spanish opinion columns. Using a contrastive analysis methodology, our investigation will identify and describe the most frequent adverbials, adjectives and verbs expressing certainty and doubt in both languages and the grammatical patterns associated with them (Hunston & Francis 2000).

The comment texts analysed in this paper have been extracted from a large comparable corpus of English and Spanish opinion columns. The source of empirical data is provided by the three most widely read British upmarket newspapers *The Times*, *The Guardian* and *The Daily Telegraph* and Spanish quality newspapers *El País*, *ABC* and *El Mundo*. The similarities and differences between both languages, revealed in this study, may be valuable in the field of translation. These authentic data could also supply the raw material for the building up of applications in the field of professional writing for the press (ESP).

References

Biber, D. and Finegan, E.T. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, 9 (1), 93-124.

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

Hunston, S. & Francis, G. 2000. *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam/Philadelphia: John Benjamins.

Affix frequency and word order in the history of English: Parsing Middle English morphosyntax

Hagen Peukert
University of Hamburg

The vast majority of the investigated languages so far display suffixation as the most predominant affixation type (Cysouw 2006; Greenberg 1957; Sapir 1921). This is true for head-final and head-initial languages. Still, prefixation is much more common in VO-languages than in OV-languages (Stump 2001: 708). A clear explanation for this phenomenon is still disputed. Givón (1979) Hawkins and Cutler (1988), Bybee et al. (1990), and Hall (1992) favor different explanations. The appropriateness of these explanations for each language remains an individual case (see Hawkins and Cutler 1988: 288).

The history of the English language is particularly interesting in that respect because it may portray a case in point allowing us to better understand the specifics and involved processes of the affix preference of OV- and VO-languages. In Early Middle English the lose word order also showing OV constructions changed to a strict VO pattern (Trips 2001). Free word order in Old/Middle English could mark a transition phase from a head-final language to a head-initial language. Provided that English really meets both criteria, that is, first prefixation is less common in OV-languages and, second, English changed from an OV to a VO-language, we may hypothesize that the number of prefixes increased throughout the history of English either in absolute terms or relative to its inventory of suffixes or both. Observing this development in detail on concrete corpus data will offer alternative explanations or will support existing explanations for the typological phenomenon of the relation between affix distribution and word order.

The architecture of a morphological analyzer and first results of the study will be presented. The software program is applied to the different time stages of the Penn Parsed Corpora of Historical English (PPCME2, PPCMBE). For each period, all affixes are stripped, counted, and categorized for word class. Comparing the frequencies of the involved word types at different points in time will allow us to make reasonable claims about the productivity of each affix distinguishing between derivational and inflectional variants. The absolute and relative numbers of all affix types can be cross-correlated with occurrences of OV and VO patterns at a given time.

References

Bybee, J. L., W. Pagliuca, and R. D. Perkins (1990), 'On the asymmetries in the affixation of grammatical material', in W. Croft, K. Denning, and S. Kemmer (eds.), *Studies in typology and diachrony: Papers presented to Joseph H. Greenberg on his 75th birthday*; Amsterdam: Benjamins, 1-42.

Cysouw, M. (2006), 'The asymmetry of affixation', in H.-M. Gärtner, et al. (eds.), *Puzzles for Krifka*: Online Publication.

Givón, T. (1979), *On understanding grammar*, New York: Academic Press.

Greenberg, J. H. (1957), *Essays in Linguistics*, Chicago: University of Chicago Press.

Hall, C. J. (1992), *Morphology and mind*, London: Routledge.

Hawkins, J. A. and A. Cutler (1988), 'Psycholinguistic factors in morphological asymmetry', in J. A. Hawkins (ed.), *Explaining Language Universals*; Oxford: Blackwell, 280-317.

Sapir, E. (1921), *Language: An Introduction to the Study of Speech*, New York: Hacourt, Brace & Co.

Stump, G. T. (2001), 'Affix position', in A. Burkhardt, H. Steger, and H. E. Wiegand (eds.), *Language Typology and Language Universals*, Handbücher zur Sprach- und Kommunikationswissenschaft; Berlin: Walter de Gruyter, 708-14.

Trips, C. (2001), *From OV to VO in Early Middle English*, Amsterdam: Benjamins.

Annotation by transformation: Advantages of TBL for tagging spoken ELF data

Michael Radeka, Ruth Osimk
University of Vienna

In this paper, we will demonstrate how the advantages of Transformation based learning (TBL) (Brill 1995) can be used to increase automatic tagging accuracy and thus to reduce the manual annotation cost.

It is commonly acknowledged that accuracies of part-of-speech (POS) taggers decrease dramatically when they are applied to different types of data than those they were originally trained on (e.g. Giesbrecht & Evert 2009). This holds especially true when systems trained on written data are applied to transcribed spoken material. In our case, the challenge is adding POS annotation to the naturally occurring, oral data of English as a lingua franca (ELF) captured in the Vienna-Oxford International Corpus of English (VOICE).

The application of the Penn-Treebank trained TreeTagger (Schmid 1994) on VOICE data serves as a valuable starting point for POS annotation (cf. Osimk forthc.). However, more satisfactory results could probably be achieved when training TreeTagger on manually annotated sections of spoken data. Building such a tagged corpus through a recurring process of automatic annotation and manual correction (e.g. Nøklestad & Søfteland 2007) is a common tagging method. However, this method involves great manual effort in order to attain the desirable amount of annotated training data. Manual annotation can provide a lot of insight into the data (Johansson 2004) and thus has the potential to increase the efficiency of the manual annotation process. However, a major disadvantage of this approach is that information about the manual corrections has no direct impact on further automatic tagging. One way of avoiding this disadvantage is to use the information gained from manual correction by applying parallel TBL (Radeka 2009).

TBL has a number of advantages which we consider helpful for efficient corpus annotation. Firstly, TBL is highly flexible as it is able to learn rules from the output of almost any initial classification model, and can therefore be used as an error-corrector for different baseline models (e.g. Ma et al. 2000, Avinesh & Gali 2007, Guanglu et al. 2008, Wu et al. 2004). Secondly, TBL rules open the data for the linguistic interpretation concerning the nature of tagging errors, error-types, their position and cause (Elming 2006). As a result, TBL rules allow the annotator to supervise and modify the generated rules or to create new ones. The interpretable rules of TBL are likely to provide systematic insight into some important aspects of VOICE data and the nature of spoken ELF in general.

References

Avinesh, P.; Gali, K. 2007. “Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning”. *Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL)*, 21-24.

Brill, E. 1995. “Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging”. *Computational Linguistics* 21, 543-565.

Elming, J. 2006. “Transformation-based correction of rule-based MT”. *11th Annual Conference of the European Association for Machine Translation*, Oslo, Norway, 219-226.

Giesbrecht, E.; Evert, S. 2009. “Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus.” *Web as Corpus Workshop (WAC5),* 27.

Guanglu, S. et al. 2008. “Chinese Part-of-speech Tagging Based on Fusion Model”. *Proceedings of The Eleventh Joint Conference on Information Science .* Shenzhen, China.

Johansson, S. 2004. “Corpus linguistics – past, present, future: A view from Oslo”. In Junsaku Nakamura; Inoue, Nagayuki; Tabata, Tomoji (eds.) *English corpora under Japanese eyes.* Amsterdam: Rodopi, 3-24.

Loftsson, H. et al. 2010. “Developing a PoS-tagged corpus using existing tools”. *Proceedings of ‘Creation and use of basic lexical resources for less-resourced languages’ (Workshop at the 7th International Conference on Language Resources and Evaluation, LREC 2010).* Valetta, Malta.

Ma, Q. et al. 2000. 2000. "Hybrid Neuro and Rule-Based Part of Speech Taggers". *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, 509-515.

Nøklestad, A.; Søfteland, A. 2007. "Tagging a Norwegian Speech Corpus." In Nivre, Joakim et al. (eds.) *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007.* Tartu: University of Tartu, 245-248.

Osimk, R. forthc. "Evaluating the applicability of existing POS practices for a corpus of English as a lingua franca (VOICE)". In Hoffmann, Sebastian; Rayson, Paul; Leech, Geoffrey (eds.). *Corpus linguistics and variation in English: Focus on Nonnative Englishes (Proceedings of ICAME 30).* Helsinki: VARIENG.

Radeka, M. 2009. "Paralleles transformationsbasiertes Lernen: Kombination von Regelmengen mit einem korpusbasierten Selektionsverfahren". Magisterarbeit. Ruprecht-Karls-Universität Heidelberg.

Schmid, H. 1994."Probabilistic Part-of-Speech Tagging Using Decision Trees." *Proceedings of the International Conference on New Methods in Language Processing,* Manchester, UK, 44-49.

Wu, D., Ngai, G.; Carpuat, M. 2004. "Raising the Bar: Stacked Conservative Error Correction Beyond Boosting". *Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal, 21-24.

Local grammar translation equivalents and contrastive linguistics

Renate Reichardt
University of Birmingham

The premise of this paper is that the local grammar of words, specifically their valency complements, guide the choice of translation equivalents. The research is inspired by Firth (1957:11) and Sinclair (1991:110) who claimed that meaning depends on the environment in which a word occurs. Valency theory, the property of a word to combine with or demand a certain number of elements in forming larger units (Emons 1974:34), is consequently utilised in the exploration of the grammar-lexis continuum.

A corpus linguistic approach was chosen to identify translation equivalents and their grammatical patterns. The corpus used for this investigation is EuroParl, consisting of European Parliament Proceedings. It therefore represents a specific language domain. However, this is not seen as a disadvantage, since a) "representativeness of a corpus must be regarded largely as an act of faith" (Leech 1991:27) and b) there is always a great degree of freedom in translation (Kenny 2005:162).

For exemplification the languages English and German are investigated, but the approach is also suitable for a wide range of languages to highlight structural similarities and differences relating to word meaning. The case study presented examines the polysemous verb CONSIDER and its German translation equivalents.

Frequency analysis was found to be useful in showing that there is interplay between meaning, i.e. the chosen translation equivalent, and valency patterns. All the patterns investigated showed, to some degree, preference for certain translations. Valency patterns are thus a useful indicator of likely meaning.

On the other hand, the contrastive analysis using a parallel corpus has also shown that there is great overall flexibility in the choice of translation equivalents. This illustrates that what is often considered as a straightforward rule-based construction process is much more flexible and unpredictable.

The utilization of corpus analysis and valency theory is found to contribute towards our understanding of cross-language variation and offers new insights into language creativity and the retention of meaning.

References

Emons, R. (1974) *Valenzen englischer Prädikatsverben* Tübingen: Max Niemeyer Verlag

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

Kenny, D. (2005) 'Parallel Corpora and Translation Studies' in Barnbrook, G., Danielsson, P., Mahlberg, M. (eds) Meaningful Texts – The Extraction of Semantic Information from Monolingual and Multilingual Corpora London: Continuum

Leech, G. (1991) 'The state of the art in corpus linguistics' in Aijmer, K., Altenberg, B. (eds) English Corpus Linguistics: Studies in Honour of Jan Svartvik London: Longman

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

A finer definition of neology in English: a corpus-based investigation

Antoinette Renouf
Birmingham City University

It has been shown (Renouf, 1993a) that lexical and semantic neology can be identified in a text corpus at surface level by automatic means. According to hypotheses empirically tested and upheld, a lexical neologism is usefully often a lexical item which occurs for the first time in a large corpus, especially one of journalistic text; and which can thus be found automatically by comparing each word in a chronological stream of fresh corpus data with a baseline index, such that each previously unseen item is deemed to be a candidate neologism. Meanwhile, a semantic neologism, or new sense of an existing word, has been shown to be realised in part by the change in the word's collocational environment (Renouf, 1993b).

However, the definitions of lexical and semantic neology which are implied above are of a particular kind, dictated by the particular linguistic approach, and the goal of finding an automated means of identification. A different focus comes into play when potential users of the analysed data are taken into consideration. Linguists, language teachers, lexicographers, translators and terminologists each have different interests in and uses for neologisms, and in this context, finer and more differentiated definitions of neology are required. For some applications, any neologistic use is of interest, while for more specific professional applications, the degree of establishment of the word in the relevant lexicon or sub-lexicon is vital.

This paper will seek to address this issue by newly combining a series of linguistic criteria and lexical-statistical measures previously used successfully to identify the changing status of a neologism in a corpus across time, tracking the 'life cycle' of a word (Renouf, 2007), from its first appearance, through the stages of its assimilation into mainstream language.

The study will be based on a corpus of one billion words of UK mainstream newspaper text covering the period 1989-2010. Comparative reference will also be made to more specialist textual domains where necessary.

References

1993a Renouf 'A Word in Time: first findings from dynamic corpus investigation' in English Language Corpora: Design, Analysis and Exploitation, eds. Aarts, Jan, de Haan, Pieter, and Nelleke Oostdijk, Rodopi, Amsterdam, pp. 279-288.

1993b Renouf 'Making Sense of Text: Automated Approaches to Meaning Extraction', in Proceedings of 17th International Online Information Meeting, 7-9 December 1993, pp. 77-86.

2007 Renouf Tracing lexical productivity and creativity in the British media: The Chavs and the Chav-nots?, in Judith Munat (ed.) Lexical Creativity, Texts and Contexts. Amsterdam: John Benjamins Publishing Company. 61-89

On pronominalization and the rise of the propword *one*: Stig Johansson and English historical linguistics

Matti Rissanen
University of Helsinki

It is perhaps not generally known that Stig Johansson started his scholarly career as a historical linguist. The title of his doctoral dissertation, accepted at Indiana University in 1968, is *Studies in the History and Development of the English Language.* The study is divided into nine chapters discussing the development of various aspects of the history of English syntax, from Old to Modern English. The topics include word order in various clause types, the origin and development of the auxiliary *do,* pronominalization and deletion, and the origin and development of the prop-word *one.*

Even in this early study, Stig Johansson's mastery in seeing structural similarities in different syntactic developments and his ability to combine solid textual analysis with more theoretical considerations is obvious. He successfully links the emphatic and non-emphatic uses of the auxiliary *do* with the deletion of elements and increasing pronominalization.

The purpose of the present paper is to elaborate on the findings which make Stig Johansson's analysis and conclusions valuable even for the historical syntacticians of the 21st century. The role of corpora in further study of the topics included in the dissertation will also be discussed. Particular attention will be called to the use of *do* with inchoative verbs and to the different syntactic constructions in which the propword *one* occurred in early English.

References

Johansson, Stig. 1968. *Studies in the History and Development of the English Language.* Diss. Indiana University. Bloomington, Indiana.

Raumolin-Brunberg, Helena and Arja Nurmi. 1997. Dummies on the Move: Prop-ONE and Affirmative DO in the 17th Century. In: Terttu Nevalainen and Leena Kahlas-Tarkka (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen.* Helsinki: Société Néophilologique.

Rissanen, Matti. 1999. Syntax. In: Roger Lass (ed.), *The Cambridge History of the English Language*, Vol. III, *Early Modern English 1476-1776.* Cambridge: CUP. 187-331.

Rissanen, Matti. 2004. Otto Jespersen and the uses of *one. In:* H.-M. Lee and Y.-B. Park (eds.), *Otto Jespersen: Festschrift for the 80th Birthday of Professor Sung-Sik Cho.* Seoul: Hankook Munhwasa. 523-536.

A comparative analysis of an invariant tag in six varieties of English, *eh*?

Melanie Roethlisberger
University of Zurich

The invariant tag (InvT) *eh?* (as in *That's the one, eh?*) is a linguistic feature distinctive of several varieties of English. Invariant tags, unlike canonical question tags (e.g. *That's the one,* ***isn't it?***), do not change their form and can be used in all grammatical contexts without being constrained by the syntactic-semantic properties of the preceding proposition (Andersen 2001: 104). Previous studies of *eh?* have mainly focused on one language variety (e.g. Meyerhoff 1994; Stubbe & Holmes 1995; Gold 2008; Gold & Tremblay 2006). Avis (1972) carried out a comparative study *of eh?* in British, American and Canadian English, but the first quantitative study of InvT from a corpus-based cross-linguistic perspective has only recently been conducted by Columbus (2009, 2010), who compared InvT in New Zealand, British and Indian English, and also Singapore and Hong Kong English. The aim of my study is thus to contribute to previous work by offering a qualitative analysis of a wider range of corpora which looks into other (lesser-known) varieties of English too. More precisely, the aim is to draw a comparative analysis of the functional distribution of *eh?* in six varieties of English, namely Scottish English, Channel Island English, Canadian English, New Zealand English, Fiji English and Maltese English.

The corpora consist of spoken data, since *eh?* is a linguistic feature characteristic of speech but rarely found in written texts (see Brinton 1996: 33). The material under scrutiny comes from the *Scottish Corpus of Text and Speech* (1945-2007), and spoken private conversation transcripts from the *International Corpus of English*: ICE-Canada (1989-2009), ICE-New Zealand (early 1990s) and ICE-Fiji (under compilation). The total estimate is of 367 files, 1,210,803 words (excluding corpora files still under compilation). For both Channel Island English and Maltese English, I rely on secondary literature.

Thus, in this paper I will investigate the full array of materials in order to shed light on whether and to what extent *eh?* is a universally InvT, yet serving different functions within the discourse of a specific variety, as a pilot study seems to suggest; for instance, *eh?* is used to ask for repetition, ask for opinion, and to maintain the narrative in Channel Island English, while being more widely applied in Canadian English (see Gold & Tremblay 2006: 249). My findings will show that cross-varietal comparisons of discourse markers are a prolific field of research which deserves more attention than has been paid hitherto.

References

Andersen, Gisle. 2001. *Pragmatic Markers and Sociolinguistic Variation. A Relevance-Theoretic Approach to the Language of Adolescents* (Pragmatics & Beyond New Series 84). Amsterdam/Philadelphia: John Benjamins.

Avis, Walter. 1972. So eh? is Canadian, eh? *Canadian Journal of Linguistics* 17:2, 89-104.

Brinton, Laurel J. 1996. *Pragmatic Markers in English: Grammaticalization and Discourse Functions*. Berlin/New York: Mouton de Gruyter.

Columbus, Georgie. 2009. A corpus-based analysis of invariant tags in five varieties of English. In: Antoinette Renouf and Andrew Kehoe (eds.). *Corpus Linguistics: Refinements and Reassessments* (Language and Computers – Studies in Practical Linguistics 69). Amsterdam/New York: Rodopi, 401-414.

Columbus, Georgie. 2010. A comparative analysis of invariant tags in three varieties of English. *English World-Wide* 31:3, 288-310.

Gold, Elaine. 2008. Which *eh* is the Canadian *eh?*. *Toronto Working Papers in Linguistics* 27, 73-85.

Gold, Elaine and Mireille Tremblay. 2006. *Eh?* and *Hein?*: Discourse Particles or National Icons? *Canadian Journal of Linguistics/Revue canadienne de linguistique* 51:2/3, 247-263.

Meyerhoff, Miriam. 1994. Sounds Pretty Ethnic, eh?: A Pragmatic particle in New Zealand English. *Language in Society* 23:3, 367-388.

Stubbe, Maria and Janet Holmes. 1995. *You know*, *eh* and other 'exasperating expressions': an analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English. *Language and communication* 15, 63-88.

The influence of lexical aspect on learner aspect (mis)use: a corpus-based comparison of learner and native writing

Svetla Rogatcheva
Justus-Liebig Universität Giessen

Studies on the second-language acquisition of tense-aspect morphology focus on the claim that in the early stages of the acquisition process, the development of tense-aspect inflections is universally influenced by the inherent lexical aspect of the verbs these inflections are attached to – a theory known as the Aspect Hypothesis (AH) (cf. Andersen and Shirai 1996). Most research on the AH has so far focused on beginners, using elicitation tasks in experimental settings.

The present report is a corpus-based pilot study which examines the claims of the Aspect Hypothesis (AH) in terms of the correlation between inherent lexical aspect and advanced EFL learners' use of aspect in English – in particular, the use and misuse of the progressive and the perfect. The corpus data for the analysis is based on the Bulgarian and German components of the International Corpus of Learner English (ICLE) and four native corpora of novice and expert writing (Granger et al. 2009). The learner components have been selected in view of the typological and functional differences between the aspect systems in German and Bulgarian as native languages and English as a foreign language, whereas the native corpora represent British and American novice and expert writing. The corpora have been tagged for parts of speech with Wmatrix (Rayson 2007) and the progressive and perfect verb phrases have been extracted manually and marked for lexical aspect.

A preliminary cross-category analysis (cf. Bardovi-Harlig 1999) of the lexical verb types of the progressive and the perfect in the learner and native corpus data shows that association between telicity and verb marking varies across the learner and native corpora and seems to be stronger for advanced German and Bulgarian EFL learners in general and with regard to misuse in particular, since e.g. a substantial part of the erroneous uses of the progressive are due to inappropriate overextension of the progressive to inherently stative verbs.

References

Andersen, R. W., & Shirai, Y. (1996). The Primacy of Aspect in First and Second Language Acquisition: The Pidgin-Creole Connection. In W. C. Ritchie, & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 527–570). SanDiego: Academic Press.

Bardovi-Harlig, K. (1999). From Morpheme Studies to Temporal Semantics: Tense-Aspect Research in SLA. *Studies in Second Language Acquisition, 21*(3), 341–382.

Granger, S. et al. (2009). *International corpus of Learner English* (Version 2.). Louvain-la-Neuve: UCL Presses Univ. de Louvain.

Rayson, P. (2007). *Wmatrix: a web-based corpus processing environment*. Lancaster: Computing Department, Lancaster University, from http://www.comp.lancs.ac.uk/ucrel/wmatrix/.

The type *cannot help* V *-ing* and its rivals in Modern English

Günter Rohdenburg
University of Paderborn

This paper charts two local and intercon-nected constructional changes that have occurred over the last four centuries: the replacement – in the 17th and 18th centuries – of the type *cannot choose but* + verb stem by *cannot help* V *-ing*, which in turn has been giving way to *cannot help but* + verb stem since the early 19th century.

The first citation of the type *cannot help but* + verb stem in OED2 dates from 1894 and represents a British source. It can be shown, however, that the type goes back to the 1790s and that it constitutes an American-led innovation. It is suggested that it may be a hybrid between the submerged type *cannot choose but* + verb stem and its replacement *cannot help* V *-ing*. While the type *cannot help but* + verb stem has by now turned into an important minority variant in BrE, it still is distinctly more common and versatile in AmE.

Assuming that the three constructions are very similar in their semantics, an attempt is made to discover any contextual factors distinguishing between them during at least part of their co-existence. The constraints discussed include the following:

1) the tendency for the perfect tense to be attracted to *cannot help but* + verb stem
2) the relative preference shown by *cannot choose but* + verb stem and *cannot help but* + verb stem for inanimate subjects
3) the special affinity of *cannot help but* +verb stem for contexts such as (i), where some element has been extracted out of an embedded clause by relativization (or some other device)

(i) These are some of the myths that the book cannot help but explode.

It is noted that the third tendency runs counter to the Domain Minimization Principle as advanced by Hawkins (1999:263).

The database used for this paper consists of (several years of) British and American newspapers from the 1990s, and a smaller collection of historical corpora provided by Chadwyck-Healey and the Gutenberg project.

Reference

Hawkins, John A. (1999).Processing complexity and filler-gap dependencies across grammars. Language 75. 244-285.

‘New’ themes and cohesion: multiple themes, information structure and cohesion in learner language

Sylvi Rørvik
Hedmark University College

This paper presents some preliminary results from a study of multiple themes, information structure and cohesion in learner language. The material comprises three main text categories: novice-writer (i.e. non-expert) texts written in English by Norwegian advanced learners, and published newspaper texts written in English and Norwegian. The novice category contains argumentative texts taken from NICLE, the Norwegian component of the International Corpus of Learner English, and the newspaper texts are argumentative texts taken from the comments section of newspapers.

The focus of the paper is on themes (defined according to Halliday 2004: 79-81) containing ‘new’ information. A previous study (Rørvik, forthcoming) found that NICLE writers underuse ‘new’ themes, as compared to expert writers of English. The paper will attempt to answer the following questions:

1) Is there any correlation between information status and the type of construction found in thematic position? One might, for instance, hypothesize that the expert writers use more marked themes in the form of circumstances/adverbials, which one might expect to contain ‘new’ information, and that these types of marked themes are underused by the NICLE writers.
2) Given that the NICLE texts contain fewer ‘new’ themes than the English expert-writer texts, what is the information status of multiple themes containing a textual theme (i.e. a conjunctive Adjunct or a conjunction, cf. Halliday 2004: 79)? One might expect the use of connectors, for instance, to increase with an increased proportion of themes containing ‘new’ information, in order to maintain cohesion in a text. Previous studies of Norwegian-produced learner English (e.g. Drew 1998: 126; Hasselgård 2009: 127) have shown that Norwegian learners tend to overuse connectors, like their Swedish peers (Boström Aronsson 2005: 75), but with the underuse of ‘new’ themes in the NICLE material, there must be some other reason for this higher proportion of connectors in the learners’ texts.

References

Boström Aronsson, M. 2005. *Themes in Swedish Advanced Learners’ Written English.* PhD dissertation, Göteborg University.

Drew, I. 1998. *Future Teachers of English: A Study of Competence in the Teaching of Writing.* Kristiansand: Høyskoleforlaget.

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

Hasselgård, H. 2009. “Thematic choice and expressions of stance in English argumentative texts by Norwegian learners.” In K. Ajmer (ed.), *Corpora and Language Teaching.* Amsterdam: Benjamins, 121-139.

Rørvik, S. (forthcoming) “Thematic progression in learner language.” In S. Hoffmann, P. Rayson and G. Leech (eds.), *Corpus linguistics: Looking back – moving forward.* Amsterdam: Rodopi.

CorTrad search features and translation studies: a pilot study on colours, clothing and food domains

Diana Santos, University of Oslo
Stella E.O. Tagnin, Elisa Duarte Teixeira, Universidade de São Paulo

In this paper we present yet another parallel corpus for the Portuguese-English language pair, CorTrad, touching upon its technical infrastructure, the specially developed features of the interface, and presenting some pilot studies on the colour, clothing and food domains.

CorTrad has currently three subcorpora: *Cortrad literário*, a fiction/short stories corpus with three translation stages (from a translator learner corpus to a published book); *Cortrad técnico-científico*, a Brazilian cookbook translated into English by Brazilian translators and then revised by a native American English speaker in two versions) and *Cortrad jornalístico*, featuring a set of 1,072 scientific magazine articles translated into English in Brazil. CorTrad is thus innovative in both content and setup.

The work done in CorTrad is the logical continuation of previous work deployed in COMPARA both in terms of expanding and improving the DISPARA system and using our experience of syntactical annotation of Portuguese with PALAVRAS (Bick 2000) and of English with CLAWS (Rayson & Garside 1998). COMPARA and CorTrad are a follow-up of the AC/DC project and in fact it is convenient for expository purposes to refer to them as AC/DC cluster since they share, for the Portuguese part, most of the processing, realizing in practice Stig Johansson's model for parallel corpus compilation, where the translated texts are part of a larger monolingual pool (Johansson & Hofland, 1994).

All texts have been semantically tagged for colour and clothing.

Colours are of special interest because they refer to visual properties and because they have a strong metaphorical import. We have already noted in COMPARA's literary texts that not all colour metaphors transfer across English and Portuguese.

As to clothing, although a less common dimension in general language, it has received quite a lot of attention in linguistics, and we hypothesise a high genre-discrimination potential. Furthermore as a cultural domain par excellence it is also of relevance for translation studies.

Finally, CorTrad allows us to study in detail the translation of recipes and references to food preparation and serving, which is a very specialized and culture-laden translation area.

We will thus present how these three subcorpora differ in terms of lexical texture, and how the domains of clothing and colours change in the different stages of translation, as a pilot study of how semantic fields change across translation versions.

Degrees of grammaticalization: Evidence from modal and aspectual verbs grammaticalizing at different rates

Monika Edith Schulz
University of Hamburg

This paper investigates synchronic variation in the distribution of the overt markers of past obligation (HAD TO and HAD GOT TO) and past habituality (WOULD and USED TO) in two spoken, traditional British English dialects. The data is comprised to two sub-corpora of 180,000 words each from the Freiburg Corpus of English Dialects (FRED), representing three North counties (Lancashire, Westmoreland, Yorkshire, and) and two Midlands counties (Nottinghamshire, Shropshire).

Synchronic variation is conceptualized here as different degrees of grammaticalization in the different dialect areas. Different degrees of grammaticalization are measured on the basis of two indicators: the relative frequencies of variants and the loosening of historical selection restrictions on the use of the variants, couched in terms of context-expansion in the sense of Himmelmann (2004).

Relative frequencies are a helpful indicator of grammaticalization for past obligation marking. We find variation between HAD TO and HAD GOT TO in the Midlands, where HAD GOT TO accounts for roughly 30% of all past obligation contexts. In the North, however, HAD TO is the default past obligation marker with HAD GOT TO at less than 1% of all past obligation contexts. Thus, the Midlands exhibit the more grammaticalized system, in line with findings in the area of present tense obligation, where a leap in the relative frequency of HAVE GOT TO has been identified as the crucial factor in the “success story” of its grammaticalization (Krug 2000).

In the area of past habituality the relative frequencies of WOULD and USED TO cannot be used as the sole criterion for degree of grammaticalization but have to be complemented by an investigation into the context expansion of USED TO. Historically, USED TO was restricted to combinations with human subjects and non-stative verbs (Visser 1969, Bybee et al. 1994). A multivariate analysis of the factors contributing to the choice between WOULD and USED TO, including subject animacy and verb type as historically motivated constraints, sheds light on the status of these historical selection restrictions.

A comparison of the evidence from the two indicators suggests that relative frequency is too coarse a tool to capture the more intricate differences between dialectal systems of past habituality marking. Weaker disfavoring effects of historically relevant constraints on USED TO do go hand in hand with a higher relative frequency of the marker most of the time. Equal relative frequencies of WOULD and USED TO, however, often gloss over different patternings of constraints and thus merit particular attention.

References

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

Himmelmann, Nikolaus. 2004. Lexicalization and grammaticalization: Opposite or orthogonal? In:Walter Bisang, Nikolaus P. Himmelmann, and BjörnWiemer (eds.), *What Makes Grammaticalization? A Look from its Fringes and its Components*, 21–43. Berlin and New York: Mouton de Gruyter.

Krug, Manfred. 2000. Emerging English Modals. A Corpus-based Study of Grammaticalization. Berlin and New York: Mouton de Gruyter.

Visser, Theodorus Fredericus. 1969. *An Historical Syntax of the English Language*, vol. 3.1. Leiden: Brill.

How specific is English for Academic Purposes?

Natassia Anne Schutz
Université Catholique de Louvain

The notion of English for Academic Purposes (EAP) has become the subject of much debate in recent years. While some authors defend the idea of a general academic vocabulary that cuts across disciplines (e.g. Coxhead, 2000; Granger & Paquot, 2009) others, such as Hyland and Tse (2007), highlight the specificity inherent in each discipline and question "the assumption that a single inventory can represent the vocabulary of academic discourse and be valuable to all students irrespective of their field of study" (ibid.: 238).

The objective of my study is to take the debate one step further by exploring the notions of EAP vs. English for Specific Purposes (ESP) in more detail. Like Granger & Paquot (2009) I focus on verbs, but while they exclusively focused on the verbs *shared* across disciplines, I analyze *all* the verbs found in the 3 million word *Louvain Corpus of Research Articles (LOCRA)*, an ESP corpus. My aim is to indentify the verbs which are shared across the three disciplines in the corpus (business, linguistics and medicine) vs. those which are specifically inherent in a particular discipline. The method used to extract the LOCRA verbs is the keyness method which consists in extracting all the words that "occur with unusual frequency in a given text" compared to a reference corpus (Scott, 1997). This procedure was used by Paquot (2007) to extract a general academic keyword list and proved fruitful as it brought to the fore many words typical of EAP which were not included in preceding lists (e.g. Coxhead's Academic Word List, 2000).

The results of my study show that 50% of the verbs are discipline-specific and that the other 50% are shared by 2 or 3 disciplines, making them likelier candidates for general academic vocabulary. Medicine stands out as having the highest number of discipline-specific verbs while business and linguistics prove to share a considerable number of verbs. To expand on these results, a selection of EAP and ESP verbs will be investigated phraseologically, showing that "they occur and behave in different ways across disciplines" (Hyland & Tse, 2007: 235). To ensure that important academic verbs are not missed by the keyness method, I compared the results with the top verbs in frequency-based lists, thereby demonstrating the benefit gained from complementing the keyness procedure with more traditional frequency-based lists. The conclusion will propose some implications for EAP and ESP teaching.

References

Coxhead, A. (2000). A New Academic Word List. In *TESOL Quarterly*, 34(2): 213-238.

Granger, S. & Paquot, M. (2009). In search of a General Academic vocabulary: A corpus-driven study. In Katsampoxaki-Hodgetts, K. ed(s). *Options and Practices of LSP Practitioners.* University of Crete Publications: 94-108.

Hyland, K. & Tse, P. (2007). Is there an "Academic Vocabulary"?. In *TESOL Quarterly*, 41(2): 235-253.

Paquot, M. (2007). Towards a productively-oriented academic word list, In J. Walinski, K. Kredens & S. Gozdz-Roszkowski ed(s) Frankfurt am Main, Peter Lang. *Corpora and ICT in Language Studies. PALC 2005*. Lodz Studies in Language, 13: 127-140.

Scott, M. (1997). PC analysis of keywords and key keywords. In *System*, 25 (2): 233-245.

“They have published a new cultural policy that just come out”: competing forms in spoken and written New Englishes

Elena Seoane, University of Santiago de Compostela
Cristina Suarez-Gomez, University of the Balearic Islands

This paper sets out to assess the variation found in the expression of perfect meaning in New Englishes. A preliminary analysis of three million words of spoken New Englishes in Asia (using a parallel corpus of British English as a benchmark corpus) revealed the use of different variants in contexts where Standard English requires the presence of *have*+past participle (Huddleston & Pullum 2002: 143), namely contexts expressing recent past with *just*, experiential meaning with *(n)ever*, and resultative meaning with *yet* (Suárez-Gómez & Seoane 2010; cf. also Miller 2000:327-331). While some of these variants showed a significant degree of productivity in all spoken registers (e.g. the preterite form), others were rarer and featured mainly in informal spoken interaction (e.g. the base form, as in *She never explain properly to her* and *have*+base form as in *That's the most interesting paper I've ever work on*); for this reason, the latter could be interpreted as arising from performance or transcription errors, rather than being an innovative form. In this paper we gauge the impact of this variation in written New Englishes, in order (i) to identify the differences between spoken and written New Englishes in the expression of perfect meaning, as compared to spoken and written British English (Miller 2006), and (ii) to ask to what extent the innovations found in spoken New Englishes have spread to written New Englishes. The occurrence of such forms in the written language would confirm a structural change, since they would represent consolidated variants within the perfect paradigm.

We will analyse the expression of the perfect in the above mentioned contexts in Hong Kong, Singapore, the Philippines and Indian English (plus British English) as represented in the ICE corpora. The samples contain 1,000,000 words for each variety, 600,000 words corresponding to spoken language and 400,000 to the written register.

References

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

Miller, Jim (2000). The perfect in spoken and written English, *Transactions of the Philological Society*, 98(2): 323-352.

Miller, Jim (2006). Spoken and written English. In Bas Aarts and April McMahon (eds.). *The Handbook of English Linguistics*. Oxford: Blackwell: 670-691.

Suárez-Gómez, Cristina and Elena Seoane (2010). The expression of the perfect in Asian Englishes. Paper presented at *43rd Annual Meeting of the SLE*, Vilnius.

Emergent complex prepositions: variability within the PNP construction

Adam Smith
Macquarie University

Hoffmann (2005) put the case for the existence of the complex preposition (CP) as a grammatical unit. By tracing the development of the 30 most common PNP constructions in the British National Corpus, from the Middle Ages to the 20th century, using material from the *Oxford English Dictionary* quotations database, and texts from Project Gutenberg, he made a strong argument for these constructions having become grammaticalised rather than simply being idiomatic expressions, as Huddleston and Pullum (2002) argue, amongst others.

This study builds upon Hoffmann's work by examining the recent grammaticalisation of more variable CPs. It expands the range of CP constructions under investigation by focussing on PNP sequences where the noun is constant, but the preposition can be variable. It includes categories specifically excluded by Hoffmann, such as those where the nominal element is preceded by a determiner. Data from the Corpus of Contemporary American English (COCA) is examined to draw American English comparisons with the British English data. The recent publication of the Corpus of Historical American English (COHA) has also made possible a more systematic diachronic study of the development of CPs decade by decade over the last 200 years. Examples of these developments include *in behalf of* being replaced by *on behalf of* as the preferred form, and the gradual replacement of *in respect to* and *in virtue of* by *with respect to* and *by virtue of* in American English. The same kind of gradual shift is shown with the loss of the definite article over the same period for *in place of*, and the more recent rise of the CPs *in light of* and *at risk of*, which previously took the definite article almost invariably.

Quirk et al.'s (1985) "scale of cohesiveness" is a useful template for the types of variability possible within the PNP unit, without providing a gradience as to which variations indicate the degree of grammaticalisation. Early findings indicate that the potential for adjectival premodification within a PNP sequence indicates a lesser degree of grammaticalisation than does interruption by a determiner. The study argues that the potential for variability, rather than disproving the category, as Huddleston and Pullum contend, shows that "complex preposition sequences display varying degrees of emerging constituency" (Beckner and Bybee, 2009).

References

Beckner, Clay and Joan Bybee. 2009. A Usage-Based Account of Constituency and Reanalysis. *Language Learning* 59: Suppl 1., pp. 27-46.

Hoffmann, Sebastian. 2005. *Grammaticalization and English Complex Prepositions: A Corpus-Based Study*. London & New York: Routledge.

Huddleston, Rodney and Geoffrey Pullum, 2002. *The Cambridge Grammar of English Usage*. Cambridge: Cambridge University Press.

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

Speech acts, communicative intention, and Grice's maxims in SMS social networks

Susana M. Sotillo
Montclair State University

This study investigates types of speech acts and adherence to Grice's maxims in the text messaging practices of members of five diverse SMS social networks.

Despite omissions of noun phrases, function words, and other features of formal writing conventions, SMS text messages have been shown to successfully convey the intended meaning to an addressee because in most cases the implicit communicative intention is successfully interpreted by the message's recipient. Thus, even though syntactic structures are simplified and there are violations of the strict interpretation of the Theta-criterion, the missing thematic roles assigned to NPs are recoverable. Recent studies of text messaging practices indicate that SMS text messaging language appears to constitute a particular variety of naturally occurring language characterized by structural simplifications and semantic implications, which is widely used across age cohorts and diverse educational and occupational backgrounds. In this study, text messages were analyzed in a 31,000-word SMS corpus using tools from corpus linguistics. These text messaging data were collected over a period of 14 months from five different SMS social networks that included 59 participants. The study addresses three research questions: 1.What types of speech acts are found in all five SMS social networks analyzed? 2. Do these vary according to participants' SMS network membership and age? And, 3. Which of Grice's maxims do texters in these SMS networks frequently adhere to? As recent studies that focus on the nature of text messages have shown, texters are aware that their creative use of homophones, abbreviations, typographic symbols and emoticons serve specific functions and strengthen ties among those sharing common interests and background knowledge (See Baron, 2009; Betti. 2008; Campbell & Park, 2008; Crystal, 2008; Fairon, Klein, & Paumier, 2006; Thurlow, 2003; Zarantonello, 2001). This investigation will show the types of speech acts found in each of the five SMS social networks. It will also seek to explain how communicative intention is achieved by adults and youth who form part of five different SMS social networks, and the extent to which texters successfully adhere to some or all of Grice's maxims.

The diphthongisation of ME *ū*: the spelling evidence

Gjertrud F. Stenbrenden
University of Oslo

'ME *ū*' comprises the reflexes of OE *ū* and OE *u+g*, but also of OE *ō+g* in some cases (e.g. OE *hūs* 'house', *fugol* 'bird, fowl', *plōg* 'plough'). Its diphthongisation to [ʊu] > [əu] > [ʌu] > [aʊ] is part of the 'Great Vowel Shift', whereby the close vowels were diphthongised and the non-close vowels were raised one step in the vowel space. This shift is traditionally dated c. 1400-1750. However, spelling evidence for the diphthongisation of ME *ū* is hard to come by, mainly because after the Norman Conquest, <ou>/<ow> were adopted instead of historical <u> for /u:/, and <ou>/<ow> may equally well correspond to [u:] as to diphthongs [ʊu]/[əu]. Uncontroversial spelling evidence is not to be had until the fifteenth century, when e.g. the 'Welsh Hymn', written in English but according to Welsh orthography, uses <ow> for historical *ū*, clearly corresponding to a diphthong. Some *LALME* sources, likewise from the fifteenth century, use <au>/<aw> for historical *ū*, showing advanced diphthongisation. However, a systematic exploitation of potential vowel-shift spellings in *LAEME*, *SMED* and *LALME* (Stenbrenden 2011) reveals that the shift was likely under way as early as the late thirteenth century. Evidence of the raising of ME *ō* is attested from the thirteenth century onwards, and since the reflex of ME *ō* did not merge with that of ME *ū*, diphthongisation of the latter must have been at least roughly simultaneous with the raising of ME *ō*. This paper examines the available evidence recorded in the three above-mentioned sources, arguing that detailed analysis of the orthographic systems in which irregular spellings are found may corroborate the early-diphthongisation hypothesis. This paper also seeks to identify the locus of change for the diphthongisation of ME *ū*.

References

LAEME:
Laing, M. 2007. *A Linguistic Atlas of Early Middle English*. University of Edinburgh. http://www.lel.ed.ac.uk/ihd/laeme1/laeme1.html
LALME:

McIntosh, A., M. Samuels, M. Benskin *et al.* 1986. *A Linguistic Atlas of Late Mediaeval English. Vols. I-IV.* Aberdeen: Aberdeen University Press.

SMED:

Kristensson, G. 1967. A Survey of Middle English Dialects 1290-1350. The Six Northern Counties and Lincolnshire [SMED1]. Lund: CWK Gleerup.

Kristensson, G. 1987. A Survey of Middle English Dialects 1290-1350: the West Midland Counties [SMED2]. Lund: Lund University Press.

Kristensson, G. 1995. A Survey of Middle English Dialects 1290-1350: the East Midland Counties [SMED3]. Lund: Lund University Press.

Kristensson, G. 2001. A Survey of Middle English Dialects 1290-1350: the Southern Counties. I. Vowels (except Diphthongs) [SMED4]. Lund: Lund University Press.

Kristensson, G. 2002. A Survey of Middle English Dialects 1290-1350: the Southern Counties. II. Diphthongs and Consonants [SMED5]. Lund: Lund University Press.

Stenbrenden, G.F. 2011. The Chronology and Regional Spread of Long-Vowel Changes in English, c. 1150-1500. Ph.D. dissertation, University of Oslo.

Defining selected expert writing skills in the ELF context: A comparative study of academic abstracts

Petr Sudicky
Masaryk University

While EFL education in general is still dominated by native-speaker modelling, in the field of academic and specialized writing the main focus has recently shifted towards the apprentice-expert continuum in which English functions effectively as a lingua franca. In such a context, the quality of being a model does no longer rest upon language nativeness, but on the authors' success of having their work recognized by the international expert community of their discipline. However, this success is still vastly dependent on the mastery of academic writing skills which tend to be learnt through formal education rather than acquired in a natural way. Following these assumptions, the present work-in-progress report summarizes the latest results of a comparative study of academic abstracts written in English by accomplished scholars/scientists and by students at American and Czech universities yielding insights into the issue of nativeness and the role of formal education in the context of academic writing.

The main aim of the whole project is to examine the degree of correspondence between expert and apprentice academic abstracts as regards the lexicogrammatical profile (distribution of key clusters and lexical frames) and logical argument structure (frequency and position of linking devices), and to find out to what extent the differences and similarities are determined by the native language of the authors and the formal training in academic writing they might have received throughout the course of their educational experience. The analysis exploits data from three small-scale comparable corpora which were compiled specifically for the purposes of this study, and which include abstracts submitted as parts of final theses by Czech students of Masaryk University (BA and MA level), abstract sections of the MICUPS project (University of Michigan), and model expert abstracts selected from established scientific and scholarly journals. Given the fact that the analysed examples of apprentice writing vary in a number of aspects (i.e. formal training experience, linguistic background, and academic level of the author), the research outcomes display a relatively complex set of relations among the individual variables (student/apprentice-expert continuum, linguistic transmission, etc), the assessment of which is especially relevant for EFL/ELF teaching practice.

Teen language in the virtual speech community: Building and analyzing a corpus of Internet media

Sali A. Tagliamonte
University of Toronto

The Internet is becoming the most common form of communication in contemporary life, particularly among teenagers and young adults (Baron 2008). What is happening to language in this environment of intense social networking and rapidly shifting norms and practices?

This paper highlights the findings a 200,000 word corpus from forty-eight North American women and men between 18-21. A crucial and singular characteristic is that it comprises the same individuals in interaction with their friends across four computer-based media: email (EM), instant messaging (IM) and texting on mobile phones (SMS) as well as a control sample of formal written language (TXT). Three linguistic variables were selected in order to place these media on the continuum between written and spoken language: i) short forms and abbreviations, as in (1-4); intensifier *so*, a recent innovation, as in (3,4); and iii) the longitudinal shift to future reference *going to*, as in (1,4).

(1) im *going to* go to bed *ttyl* BYE (IM/r00)
(2) I *dono*, lets ask him *tmrw*. (SMS/s00)
(3) yeah *omg* im already *SO* TIRED (IM/m00)
(4) Which game *u gonna* see? *will* u see Michael Phillips, he is *sooooo* cool. (EM/x09)

Among the most striking preliminary findings is that these new media mirror existing sociolinguistic trends: e.g. females favour standard forms while males eschew them. The media differences offer important insights into the nature of Internet language: short forms are surprisingly rare (1.7% overall) but they are most frequent in IM, while EM lags far behind, patterning with TXT. Intensification rates hover around 30% across media parallel with other North American varieties; however, the frequency of incoming *so* is very high, particularly in IM. The rate of *going to* also parallels earlier studies; however, the variant forms clearly distinguish the media, with IM in the lead for *gonna*, TXT a bastion of older variants, e.g. *shall,* and EM with the widest range of forms.

Taken together these and other findings reveal teen language in virtual speech communities reflects the dynamic as well as the stable linguistic situation of the broader context (see also Herring 2003). While the contrast across media exposes a continuum of forms and patterns, the comparison of individuals across media confirms that there is no degeneration of grammar. Instead, these young people are fluidly navigating a complex range of new written registers – from formal traditional TXT to funky playful SMS – and they command them all.

Which came first, *try to* or *try and*? A chicken-and-egg story

Gunnel Tottie, University of Zurich
Sebastian Hoffmann, University of Trier

The verb *try* is unique in English in that it can take a verb complement with either *to* or *and,* without a change in meaning, as in (1) and (2):

(1) We will *try to help* you
(2) We will *try and help* you

Why is this the case? It has been suggested by Rohdenburg (2003) that the reason for using *try and* can be found in *horror aequi,* i.e. avoidance of repetition of *to,* as in (3), and substituting *and* as in (4):

(3) We want *to try to* help you
(4) We want *to try and* help you

Hommerberg & Tottie (2007) have shown that *try and* is the dominant construction in speech, with or without a preceding *to.* The *horror aequi* effect does indeed operate in Present-Day British English, especially in writing, but the effect is weak in speech; *try and* is pervasive even when no *to* precedes. This gives rise to the question of the historical validity of the *horror aequi* effect. In his study of verb complementation in 18th and 19th century English, Vosberg (2006) assumes that *try to* is the earlier construction and that *horror aequi* has given rise to the *try and*-construction:

> Dabei stellt das Syntagma *to try* den Ursprung und damit die für die Ausbreitung der Struktur *and*+Verbstamm günstigste Umgebung dar (und *and try* die ungünstigste) (2006:217)
>
> 'The syntagm *to try* is the original locus as well as the most favorable one for the spread of the structure *and* + base form of verb (and *and try* the least favorable one)'

However, although OED lists *try to*+V before *try and*+V (s.v. 16a, b) under the headword *try,* a careful historical corpus study shows that *try and*+V is the earlier construction. Using the OED quotations as a database and searching Early English Books Online (EEBO) enables us to follow the slow emergence of *try to*+V, and to trace it to the semantic development of *try* from meaning 'sort,' 'test,' to 'attempt,' which is a pre-requisite for the development of *try and*+V constructions.

References

Hommerberg, Charlotte, and Tottie, Gunnel. 2007. *Try to* or *try and? ICAME Journal* 31:43-62.

Rohdenburg, Günter. 2003. Cognitive complexity and *horror aequi* as factors determining the use of interrogative clause linkers in English. In G. Rohdenburg and B. Mondorf (eds). *Determinants of Grammatical Variation.* 205–249. Berlin: Mouton de Gruyter.

Vosberg, Uwe. 2006. *Die grosse Komplementverschiebung. Aussersemantische Einflüsse auf die Entwicklung satzwertiger Ergänzungen im Neuenglischen.* Tübingen: Gunter Narr.

Non-standard English in Popular Culture: the revenge of English from below

Joe Trotta
University of Gotenburg

In this talk, I present three brief corpus-based case studies of controversial or questionable English grammar in Popular Culture, as exemplified in the following:

1 a. *Winston tastes good like a cigarette should.* (Winston cigarette ad, 1954)
b. *The funnest ipod ever* (from Apple's ipod advertising campaign, 2008)
c. i*'m lovin' it.* (The current McDonalds ad slogan)

With such illustrations as a platform, I look at the grammatical construction in question, discuss the pertinent linguistic issues and, using the Brown Family of corpora, the COCA corpus, the BNC and the ANC, I examine what the corpora can tell us about these usages in present-day English. In each of these cases, I present data concerning the frequency of the forms in question, both before and after their infamous use in Popular Culture. In some cases, particularly the Winston ad, it appears that the dissemination of the non-standard form through the advertising campaign may have provided a tipping point for a usage shift.

In addition, I show that underlying such usages, there are broader cultural and social issues which must be understood and, in the process, I sketch out some of the most central questions about the relationship between Popular Culture and English. I also illustrate the usefulness of Popular Culture artifacts as a basis for the study of language in general and English in particular.

The treebank.info project

Peter Uhrig, Thomas Proisl
Erlangen University

The treebank.info project aims to provide a user-friendly interface to a complex and user-definable annotation pipeline from raw text to fully syntactically parsed sentences. Queries for (partially filled) syntactic structures are intuitive and fast, even for large corpora (>100 million words). The pipeline offers a choice of tokenizers, part-of-speech taggers and syntactic parsers. As the queries are based on Stanford Dependencies (de Marneffe et al. 2006), which are in turn generated from Penn Treebank style phrase structure trees, only Penn Treebank style tokenization is available. The range of PoS taggers offered includes HunPos (TnT clone – HMM based), Stanford Tagger (maximum entropy), TreeTagger (decision trees), OpenNLP (maximum entropy), MXPOST (maximum entropy), SVMTool (support vector machines). For phrase structure parsing, the Stanford, Berkeley, OpenNLP, Charniak-Johnson, BitPar and Bikel parsers are available. In the long run, the interface might also include dependency parsers such as the Malt Parser or RelEx. Particular emphasis was put on the speed and scalability of the system, thus a document database is used to hold the corpora, annotations and metadata while a graph database allows efficient querying of dependency structures.

The following features will be shown in the demonstration:

- search for abstract (not lexically filled) syntactic structures (e.g. nouns with two premodifying adjectives, ditransitive verbs, ...)

- search for partially filled structures where lemma, wordform and/or part of speech can be specified (e.g. sentences with *death* premodified by an adjective, lemma *give* with a direct object and an indirect object which is a personal pronoun, …)
- search for "collo-items" (collocations, collostructs) with abstract or lexically filled structures which can be sorted by alphabet, frequency, or association measure (e.g. premodifying adjectives of *death*, premodifying adjectives of *resemblance* as direct object of the lemma *bear*.

We will show how the treebank.info project enables users without a background in Natural Language Processing to benefit from parsing their own corpora and being able to search their own parsed corpora as well as publicly available ones.

Reference

Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning (2006), *Generating Typed Dependency Parses from Phrase Structure Parses.* In LREC 2006.

Can articles predict the word class of the premodifier? A study of the *-ing* participle

Turo Vartiainen, University of Helsinki
Jefrey Lijffijt, Aalto University

Word classes are fundamental units of linguistic analysis. However, linguists often have differing opinions regarding the boundaries of word classes. In our study, we focus on a class of premodifiers that has received varying attention in recent linguistic literature: the premodifying -ing participle, such as the forms in *an interesting play* or *the advancing troops*. While many linguists regard both *interesting* and *advancing* in the previous examples to be adjectives (e.g. Borer 1990, Biber et al. 1999), we subscribe to the idea that the participle category should be split into two classes: adjectival participles (e.g. *an interesting man*) and verbal participles (*the advancing troops*; see Huddleston 1984, Laczkó 2001).

This splitting approach has traditionally been justified on distributional grounds: words like *interesting* behave morphosyntactically like central adjectives, whereas words like *advancing* do not. A recent study brings further justification for the split, showing that the two classes are distributed differently in different registers of language use (Vartiainen and Lijffijt, submitted).

In this study we approach the question from a different angle, concentrating on the kinds of noun phrase in which the participles occur. We will focus on two syntactic patterns: *a(n)* + *-ing participle* + *noun* and *the* + *-ing participle* + *noun*. We find that the definite article is much more commonly used with verbal participles than with adjectival participles, which we explain by functional differences between verbal and adjectival participles. More precisely, instead of a characterising function, many verbal participles have a subcategorising function, and the information expressed by the participle is presented as given or accessible with the definite article (e.g. *the underlying layers*, *the remaining chapters*).

We will conclude by studying how well the word class of the premodifying -ing participle can be predicted by the definite article and how articles can be used to improve the automated POS annotation of -ing modifiers and their semantic analysis. We will start with polysemous -ing forms like *outgoing* (e.g. *an outgoing person*, *the outgoing government*), where the definite article alone is able to disambiguate the participle's word class and meaning in ca. 90% of all cases, and then proceed to study adjectival and verbal participles more generally. We find that, in our pilot data, the definite article is a reliable predictor of the participle's word class, achieving ca. 82% precision with verbal participles.

References

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

Borer, Hagit (1990). "V + ing: It Walks like an Adjective, It Talks like an Adjective". *Linguistic Inquiry* 21: 95-103.

Huddleston, Rodney (1984). *Introduction to the Grammar of English*. Cambridge: Cambridge University Press.

Laczkó, Tibor (2001). Another Look at Participles and Adjectives in the English DP. In Miriam Butt and Tracy Holloway King (eds.). *Proceedings of the LFG01 Conference*. CSLI Publications.

Vartiainen, Turo & Jefrey Lijffijt (submitted). Premodifying -ing participles in the parsed BNC. *Proceedings of the 31st ICAME conference*.

A unit of meaning: the psycholinguistic reality of the model

Svetlana Vetchinnikova
University of Helsinki

In this paper I will show how the concept of a unit of meaning, which is essentially corpus-driven, can be further explored by juxtaposing concordance data and word association responses.

Although the psycholinguistic reality of co-occurrences observed in corpora seems to be intuitively plausible, and for example Hoey has been probing this hypothesis with the lexical priming theory, it is hard to make any definitive claims about an individual's intuition, performance or mental lexicon on the basis of a corpus which is usually no one's experience of language. The kind of data I will be using for the development of my argument is different in this respect: it consists of several Master's theses written in drafts, with drafts of each thesis compiled into separate "individual" corpora. In order to get additional, psycholinguistic insight into the usage patterns observable in one student's writing, the words used by a student in his/her thesis are also given to him/her as stimuli in word association tasks. The correlation of individual concordance data and word association responses indicates that:

1) In line with Sinclair's hypothesis that "the 'core' meaning [of a word] is the most frequent independent sense" (1987: 323), in a word association task language users indeed tend to react to such "core" meaning of a stimulus word, the one which is not delexical and is "least dependent on the cotext" (Sinclair 2004: 133), even if they themselves habitually use the word in a different pattern (*bear – forest*, vs. *bear in mind, hand – foot* vs. *on the other hand).*
2) Words which alone have a relatively independent meaning of their own tend to elicit a paradigmatic word association response in which a respondent interprets the meaning of the stimulus word (*also – plus, presumably – doubt*). In contrast, words whose meaning is incomplete without the contribution of other "accompanying" words tend to elicit syntagmatic responses (e.g. *according – to*).

3) Students report that syntagmatic responses are easier to give than paradigmatic ones.

Taken together this evidence suggests that the relevant units for the mental lexicon are units of meaning as it is meaning, not form, that seems to perform the organizing function. In terms of meaning, some words seem to be more independent than others. "Dependent" words or elements which form a unit of meaning are glued together by syntagmatic association. Hypothetically, syntagmatic association is what facilitates spreading activation but then it operates *inside* units of meaning.

References

Sinclair, J.M. (1987). Collocation: a progress report. In R. Steele & T. Treadgold (eds.), *Essays in honour of Michael Halliday,* 319-331. Amsterdam: John Benjamins.

Sinclair, J. M. (2004). *Trust the text*. London: Routledge.

Modality and the V *wh* pattern

Benet Vincent
University of Birmingham

Research using corpora has provided evidence of the interdependence of paradigmatic and syntagmatic choices, which have traditionally been considered distinct, thereby offering support to Sinclair's (2004) claim that the phrase is the 'essential building block of English' (Hunston, 2003: 58). An example of this interdependence is that the paradigmatic choice of *decide* rather than *decided* seems strongly related to the syntagmatic choice of a following *wh*-clause instead of a *that*-clause (Hunston, 2003). The association between the base form *decide* and *wh*-clause coupled with the fact that the base form is also associated with modal or modal-like language led Hunston (2010) to hypothesise that modal-like language is attracted to the V *wh* pattern (Francis et al, 1996).

However, it has yet to be demonstrated whether this observation about DECIDE reveals a systematic pattern applying to all verbs that frequently govern interrogative clauses. This paper will present the results of a study that sets out to test Hunston's hypothesis by investigating verbs that frequently appear in the V *wh* pattern in the British National Corpus (BNC), using the CQP-edition of the online interface BNCweb (Hoffman & Evert, 2008). A quantitative approach is used to establish whether there is a significant attraction between *wh*-clauses and modal or modal-like language across verbs that govern the V *wh* pattern. A more qualitative methodology is used to identify the kinds of phraseologies that emerge from the data and to establish the extent to which queries based on the infinitival operator *to* followed by V *wh* can be said to result in hits containing 'modal-like expressions' (Hunston, 2010). Implications for pedagogy will also be discussed.

References

Francis, G., Hunston, S. and Manning, E. 1996. *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins.

Hoffman, S. & Evert, S. 2008. *BNCweb (CQP-Edition)*. Online resource. Available at [http://bncweb.lancs.ac.uk/] (Accessed 5/10/2010)

Hunston, S. 2003. "Lexis, wordform and complementation pattern: a corpus study". *Functions of Language* 10: 31-60.

Hunston, S. 2010. *Corpus approaches to evaluation: Phraseology and evaluative language*. New York/London: Routledge

Sinclair, J. 2004. *Trust the text*. London: Routledge.

Semantic sequences of collocational framework in different genres

Suxiang Yang
Henan Polytechnic University

Hunston (2008) puts forward the concept of semantic sequences, which is "series of meaning elements" that occur regularly in a corpus. The sequences may consist of core word, the complementation pattern or patterns associated with that word, and it can also be observed in specific corpus using grammatical words as search terms to character the texts. Semantic sequence is useful in distinguishing different genres.

This paper applied this theory to explore the semantic sequence from another perspective – collocational framework *the * of*, which is a combination of two grammatical words. The reason for choosing "the * of" as the core (search word) is that it is the most frequency of collocational framework. We assume that *the * of* could be studied from semantic sequences aspects to show the features of different texts. The corpora used are 'w ac medicine' (1,5 million) and 'w biography' (3,5 million) in BNC written. Considering there are different running words in the two corpora, the raw frequency are normalized to a common base of 1 million words. For the semantic classifications of 'w ac medicine' corpus, Marco (2000)' classification and classification of a medical dictionary were used, and for the 'w biography' corpus, Biber et al. (1999)'s classification of nouns and verbs were adopted and at the same time some adjustments were made to adapt to the features of biography. We first observe the semantic sequences in the two corpora, and then compare their semantic sequences so as to see if semantic sequences can character and distinguish genres.

The findings show that there are different semantic sequences in the two genres. In 'ac medicine' corpus, the semantic sequences is "(event) + change/relation/study + *the* + number/diagnosis/treatment + *of* + disease/tissue", while the semantic sequences in 'w biography' corpus is "activity/simple occurrence/mental + *the* + time/place/action/state/number + *of* + proper name/life experience/address(title). Therefore, there are different semantic sequences in medical texts and biography text, and semantic sequences can be employed in distinguishing different genres.

The significances of the study are: 1) theoretically, semantic sequences containing *the * of*, which can be used to distinguish different genres; 2) In practices, especially for language teaching and learning, the students can identify, understand and master the semantic sequences of different genres, and learn the semantic schemata in different language community.

Posters

The Story element in engineering lectures: an analysis of one category of pragmatic mark-up

Sian Alsop
Coventry University

This paper examines the pragmatic function of "Story" in English-medium engineering lectures from the Engineering Lecture Corpus (ELC) (currently over 60 transcripts, see http://www.coventry.ac.uk/elc). The analysis will focus on transcripts of lectures from Coventry University in the UK, Universiti Teknologi in Malaysia and Auckland University of Technology in New Zealand. The ELC video transcripts have been marked up according to five pragmatic categories: Story, Housekeeping, Summary, Humour and Prayer. A number of these elements have been assigned attributes. For example Summary contains four attributes, relating to previews and reviews of current lectures, previews of future lectures and reviews of past lectures. Humour, which has proved the most difficult to subdivide objectively, has been assigned eleven attributes, including "joke", "irony", and "black humour"; a rich resource for cross-cultural comparison. The current ELC mark-up allows only one attribute for Story, but there is potential to subdivide this element into "personal narrative" and "work-related narrative" attributes.

Although there is an extensive literature relating to narrative theory and corpus stylistics, there has been almost no research into the use of narratives in spoken academic corpora. Lectures serve both an informational and an interpersonal purpose, in that they convey disciplinary knowledge and establish relationships between participants. This dual purpose is explored by Dyer and Cohen (2000), who look at the expression of expertise and equality through narratives in two engineering lectures. Our study examines the distribution, structure and purpose of narratives in a much larger number of lectures, and also considers these narratives from a cross-cultural perspective.

The two major corpora of spoken academic English, BASE and MICASE, are not marked up for pragmatic features, although The *MICASE Handbook* lists "pragmatic highlights" for each speech event in the corpus. This paper will also discuss the benefits and pitfalls of pragmatic mark-up, which enables us to identify and group stretches of discourse which serve a similar pragmatic function, but which is inevitably subjective, at least to a certain extent.

Investigating the problem of codifying linguistic knowledge in two translations of Shakespeare's sonnets: a corpus-based study

Flávia Azevedo
Federal University of Santa Catarina

This study deals with the problem of codifying linguistic knowledge in a parallel corpus, in other words the process of corpus annotation. The study aimed at investigating two ways of dealing with parallel corpus annotation: the identification of features defined as universals of translation (Baker, 1993) and the identification of types of translational correspondence, as defined by Thunes (1998). The parallel corpus is of a special kind, since it is made up of Shakespeare's sonnets and two distinct translations into Brazilian Portuguese. The study can be considered complex for two main reasons: first, the literature does not offer clear criteria to the identification of the features defined as universals of translation, and second, the translational correspondences involve metrics and other elements of lyrical poetry.

The initial results of the application of both models (Universals of Translation and Translational correspondences) in the corpus indicate that Thunes' model is empirically more effective when applied to classify alignment units in a parallel corpus.

Register profiling of scientific texts: Experiences in linguistic description and corpus-based methods

Sabine Bartsch, Technische Universität Darmstadt
Elke Teich, Universität des Saarlandes

Register variation is a widely studied phenomenon in English linguistics. Numerous descriptive accounts have been devoted to studies of written vs. spoken language (e.g., Biber (1988, 2006) as well as synchronic and diachronic studies (e.g., Biber (1988, 1995), Biber & Finegan (1988), Biber et al. (2007).

Since registers are said to be characterized by typical clusters of features which have a greater-than-random tendency to occur (Halliday & Martin 1993, p. 54), register analysis is inherently quantitative; and because frequency of occurrence is relative, register analysis must be comparative. In order to be able to provide quantitative results, register analysis typically must be corpus-based; and in order to be comparative, appropriate corpus designs must be adopted in order to provide a suitable basis for comparison (cf. Bartsch & Teich, to appear).

We report on the results of several studies of English scientific registers carried out in the last few years on the basis of a corpus of scientific research articles from nine disciplines (computer science, linguistics, computational linguistics, biology, bioinformatics, electrical engineering, microelectronics, mechanical engineering, computer-aided construction), altogether comprising approx. 17 million tokens. The studies are concerned with (a) the distinctive linguistic properties of scientific writing (compared to less specialized texts) (cf. Teich & Fankhauser (2010)) and (b) the distinctive linguistic properties of individual disciplines (cf. Teich & Holtz (2009), Degaetano & Teich (to appear), Bartsch et al. (submitted)). Starting with commonly employed features at the level of tokens and words (type-token ratio, field-specific vocabulary/terminology), we are working our way upwards through the levels of linguistic organisation to lexico-grammatical patterns/colligation (Thompson & Hunston 2006), but also to higher, i.e. more abstract and potentially more interesting levels, such as characteristic choices of process type (Halliday & Matthiessen, 2004), thematic structure and textual organisation such as (lexical) cohesion. Together, these features form a grid of relevant categories for the analysis of linguistic variation in highly specialized domains based on the parameters of field, tenor/attitude and mode/medium of discourse (cf. Quirk et al. (1980)).

Apart from presenting the core descriptive results of our studies, we also discuss the main computational techniques employed in corpus annotation, feature extraction and data analysis.

References

Bartsch, S. & Teich, E. (to appear). Register profiling for highly specialized domains: methods and techniques. Anglistentag 2011, Freiburg.

Bartsch, S., Teich, E. & Tragl, C. (submitted). Patterns of cohesion in informationally dense texts. Submitted to Corpus Linguistics 2011, Birmingham, UK.

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

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

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

Biber, D., & Finegan, E. (1988). Drift in 3 English Genres from the 18th to the 20th Centuries – a Multidimensional Approach. *Corpus Linguistics, Hard and Soft, 2*, 83-101.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2007). *Longman Grammar of Spoken and Written English* (7. impr.. ed.). Harlow: Longman.

Degaetano, S. & Teich, E. (to appear). The lexico-grammar of stance: an exploratory analysis of scientific texts. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGFS) 2011, Potsdam.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed./rev. by Christian M.I.M. Matthiessen. ed.). London: Arnold.

Halliday, M.A.K., & Martin, J. R. (1993). *Writing Science: Literacy and Discursive Power*. London, Washington: The Falmer Press.

Quirk R., Greenbaum S., Leech G., Svartvik J. (1980). *A Comprehensive Grammar of the English Language*. London: Longman

Teich, E. & Fankhauser, P. (2010). Exploring a corpus of scientific writing using data mining. In Gries S., S. Wulff & M. Davies (eds.), *Corpus-linguistic applications. Current studies, new directions*, Rodopi, Amsterdam and New York, pp.233-247.

Teich, E., & Holtz, M. (2009). Scientific registers in contact: an exploration of the lexico-grammatical properties of interdisciplinary discourses. *International Journal of Corpus Linguistics,* 14:4, 524-548.

Thompson, G., & Hunston, S. (2006). *System and corpus: exploring connections*. London: Equinox.

Multilingual contrastive analysis of document titles in UN resolutions

Jing-Xiang Cao
Dalian University of Technology

Corpora have become essential for knowledge acquisition for Machine Translation. Translation rules can be acquired by contrastive analysis of the translation equivalents. Corpus-based Noun Phrase (NP) decomposition and cross-linguistic contrast can produce valuable translation rules.

All the resolutions passed by the 62nd Session of the United Nations General Assembly were downloaded from the UN official document center. The six language versions are aligned by the paragraph and a total of 381 document titles were extracted. The documents titles are typical complex NPs, and thus can be used to contrast the composition of NPs of different languages. Limited by the author's command of the languages, only four languages, English, Chinese, French and Russian were contrasted.

First, the head words are extracted and aligned to see their translation equivalents. Then their valencies (modifiers) are categorized into purpose, place, time, measure and other semantic fields, which are structured by means of word order, preposition or declensions. After the thorough decomposition of the extracted document titles, translation rules are summarized in the hope to facilitate Multilingual Machine Translation.

CEPhiT: Texts 'Concerning Human Understanding'

Begoña Crespo, Isabel Moskowich
University of A Coruña

The aim of this poster is to present the latest advances in the compilation of the Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing (CC).

CEPhiT (Corpus of English Philosophical Texts) is the second sub-corpus in the whole project being developed by MuStE (Research Group for Multidimensional Corpus-based Studies in English) in the University of A Coruña (Spain). This poster will present *CEPhiT* as a highly recommended tool for the historical study of English Scientific Writing. It contains philosophical texts written in English between 1700 and 1900 and, in that sense, it complements *CETA*. Since the former contains samples from what UNESCO has classified as Humanities and the latter belongs to Natural Sciences, the completion of *CEPhiT*, following the same compilation principles as *CETA*, allows for contrastive linguistic analyses.

Similarly, to facilitate sociolinguistic research, we have included, when possible, some personal details about the authors and the samples in a separate XML file.

As with the rest of text, the *Coruña Corpus Tool* (*CCT*) will make possible to retrieve information from the compiled documents, both texts and metadata files.

Directive phrasal units in business meetings

Dorothea Halbe
University of Trier

Corpus linguistics has made the search for phraseological aspects of language much easier and this has resulted in the empirical finding that language users often draw on more or less prefixed patterns when producing language and that these conventionalised phrasal units allow hearers to decode messages faster. In her corpus study Aijmer (1996) established that directives are not formed randomly, but follow specific (culturally determined) patterns and thus make use of conventionalised phrasal units such as *could you, can you* (already noted by Searle 1975) but also less prototypical units such as *will you*, which because of their lesser conventionalisation allow for more ambiguity and, thus, for more negative politeness (Brown & Levinson 1987). This – apparent – giving of freedom is especially important in international and inter-organisational business settings, where roles and obligations are continuously negotiated.

Since little research has been done on how exactly directive utterances are formulated in business and which phrasal units (such as *could you, will you, we have to*) are preferred (but see Handford 2010, Koester 2006, 2010, Vine 2004), this paper investigates the frequency and usage of selected patterns in three different settings: native speaker (NS) casual conversation, NS business meetings and Business English as Lingua Franca (BELF) business meetings.

The investigation starts by showing that there is no one-to-one form-function relation as the same form can have multiple functions (e.g. *would you, can you* can be used to ask a question or to make a request) and the same function can be expressed in multiple ways (e.g. *would you, can you* can both realize requests). The analysis continues by showing how native speakers in different settings prefer different strategies to express directives, which results in *can you* being more common in casual conversation and *we have to* in business settings (both NS and BELF). Within these conventionalized patterns, moreover, some units are used more often to formulate requests (*can you*) and some are less conventionalized and more ambiguous (*will you*) and some are used more often to express directives in business settings and rarely in casual settings (*would you*). Finally, the implications of these differences are discussed and suggestions for second language acquisition given.

References

Aijmer, K (1996): *Conversational Routines in English*. London: Pearson.

Brown, P & Levinson, S (1987): *Politeness: Some Universals in Language Usage*. Cambr.: CUP.

Handford, M (2010): *The Language of Business Meetings*. Cambridge: CUP.

Koester, A (2006): *Investigating Workplace Discourse*. London: Routledge.

Koester, A (2010): *Workplace Discourse*. London: Continuum.

Searle, J (1975). Indirect speech acts. In: Cole, P & Morgan, J (eds): *Syntax and Semantics. Vol. 3: Speech Acts*. New York: Academic Press.

Vine, B (2004): *Getting Thing Done at Work*. Amsterdam: Benjamins.

A contrastive corpus study of Chinese (L1) and English (L2) lessons

Ane He
The Hong Kong Institute of Education

This presentation reports a contrastive study of Chinese (L1) and English (L2) lessons in Hong Kong secondary schools. Aiming to explore how to take advantage of L1 academic proficiency for the benefit of L2 development, this study asked in which ways instruction of Chinese and English differed regarding content knowledge covered and levels and depth of the coverage.

Despite a demand for very high levels of English proficiency, L2 instruction in Hong Kong does not appear to be able to deliver the desired outcome with users/speakers proficient biliterately and trilingually. The problem might be traced back to the 'monolingual-oriented' language pedagogy (Cummins, 2005, p. 4). It is in this context that the contrastive study was undertaken.

A small corpus with two subcorpora was complied for the study, one containing nine Chinese lessons and the other nine English lessons with a total of 69,000 words (token). Keywords was chosen as the main analytical tool. Each of the lessons was compared with a reference corpus composed of the rest of the lessons in the sub-corpus of Chinese or English, generating a keyword list of the lesson. The Chinese keywords lists were then contrasted with those of the English to identify salient patterns of differences. Word concordances were also exploited to provide contextual information on the keywords identified.

The analysis revealed three differences, namely, the 'aboutness' of the lessons, the 'abstractness' of the instruction (as reflected in the use of metalanguage), and the 'richness in coverage' of meaning. The findings indicated that the Chinese lessons were either text-based, involving discussion of a variety of literary concepts in metalanguage, or task-based, engaging students in debating/writing about a current issue; whilst the English lessons were linguistic knowledge-based, composing of mainly discrete grammatical drills through simple classroom routines. This study argues that the three differences identified were deficiencies in the L2 classrooms; but more significantly they could also be areas where positive crosslingual transfer might be possible.

The presentation will discuss the usefulness of keywords analysis in identifying patterns of language use in the classrooms of the two cognately unrelated languages – Chinese and English; and it will also discuss a number of challenges encountered during data processing/analysis of the study.

Lexical innovation in South Asian Englishes: a corpus-based study of structural nativisation

Christopher Koch
Justus Liebig University Giessen

In terms of linguistic criteria, South Asia is often portrayed as a largely homogenous setting. In line with this belief (cf. McArthur 2002), the authors of major descriptions of individual varieties (cf. Baumgardner 1993, 1998; Nihalani et al. 2004) claim that, for the most part, their descriptions are applicable for the whole South Asian region. This view has been challenged by other researchers (cf. e.g. Gargesh 2009), who stress differences between the varieties as well as local peculiarities. Lacking large and authentic corpora, however, even most influential studies (cf. Baumgardner 1998, Nihalani et al. 2004) still relied heavily on introspection, making an objective description of unity and diversity in South Asian varieties of English impossible.

Utilising large databases which have been semi-automatically extracted from the archives of major daily newspapers in Pakistan, India and Sri Lanka (3 million words each), a re-examination of previous studies in the area of lexis is carried out, revealing their reliance on the markedness of forms (e.g. *tiffin carrier*, *co-brother/-daughter*) and their subjective character as well as largely falsifying their results. Complementing this analysis, an extensive search for yet unattested forms (e.g. *bus stand*, *traffic rule*, *upgradation*) is carried out. Since current corpus-linguistic software does not provide options to efficiently conduct such a task, semi-automatic methods are employed to extract potential candidates of lexical innovation from the source data and compare them to both a reference corpus and an online dictionary of British English (i.e. the historical input variety).

The poster presents the results from both analyses, contrasting introspective examination with corpus-based evidence as well as detailing the range and depth of integration of those lexical items that can be attested in the data. Building solely on linguistic evidence, it finally leads to the possibility of taking on an empirically based standpoint in the debate about unity and diversity in South Asian Englishes. Regarding lexical innovations, their acceptance and pervasion as markers of structural nativisation (cf. Schneider 2003, 2007), it becomes possible to infer the degree of norm development in the respective varieties of English.

References

Baumgardner, Robert J. (1993): *The English Language in Pakistan*. Karachi: Oxford University Press.

Baumgardner, Robert J. (1998): “Word-formation in Pakistani English”, *English World-Wide* 19(2), 205-246.

Gargesh, Ravinder (2009): “South Asian Englishes”, *The Handbook of World Englishes*, eds. Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson. Malden: Wiley-Blackwell. 90-113.

McArthur, Tom (2002): *The Oxford Guide to World English*. Oxford: Oxford University Press.

Nihalani, Paroo, R.K. Tongue, Priya Hosali & Jonathan Crowther (2004): *Indian and British English: A Handbook of Usage and Pronunciation.* 2nd ed. Delhi: Oxford University Press.

Schneider, Edgar W. (2003): "The dynamics of New Englishes: from identity construction to dialect birth", *Language* 79(2): 233-281.

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

The Coruña Corpus Tool: a means to an end

Inés Lareo
University of A Coruña

The Coruña Corpus Tool (*CCT*) has been developed by the Irlab (Information Retrieval Lab) and the MuStE group, both from the University of A Coruña, to explore all the possibilities the *Coruña Corpus* offers. It allows researchers to work with more than one subcorpus of the CC or to select one of them. Besides, the metadata included in each subcorpus allows the CCT to select texts applying linguistic and extralinguistic criteria. Thus researchers can compile their work-corpus following their own interests.

I will deal with the technical issues and options the CCT offers, presenting also some of the decisions that have come to constitute our editorial policy.

The Coruña Corpus Project

Isabel Moskowich-Spiegel, Begoña Crespo
University of A Coruña

The Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing (henceforth, *CC*) is a project on which the MuStE Group has been working since 2003. It has been designed as a tool for the study of language change in English scientific writing in general as well as within the different scientific disciplines. A rough definition of our corpus project would say it contains English scientific texts, other than medical, produced between 1700 and 1900.

Two ideas have triggered the project: on the one hand, the growing interest in the vernacularisation of Science in late-medieval and modern England as an understudied area and, on the other, the gradual increase in studies on genre conventions and special languages. In order to ensure fruitful linguistic analysis the selection of texts has been based on certain external parameters such as discipline classification (UNESCO classification of Sciences was the starting point for text selection), time-span (texts produced after Empiricism and the generalisation of the scientific method) and representativeness (according to the stratified sampling method proposed by Biber, 1993).

With these premises we have compiled different sub-corpora which are independent entities though sharing a similar structure, organisation and mark-up. We are compiling corpora from the Natural and Exact Sciences particularly Astronomy, Life Sciences and Chemistry, and from the Humanities (Philosophy, Linguistics and History) which facilitate comparative studies on the language of different disciplines, and the evolution of particular features of each of them, confirming the wide range of variation within academic prose.

Each subcorpus is accompanied by a corpus management tool, The *Coruña Corpus Tool* (*CCT*), an Information Retrieval system where the indexed textual repository is the set of compiled documents that constitutes the *Coruña Corpus* or any of its parts.

It works as most concordance programmes but offers some special features adapted to the characteristics of *CC* (possibility to search old-fashioned characters, tags in texts or in metadata files, for instance)

Apart from the sample selected, that is, the file with the document encoded in XML format, a metadata file containing information about the author's life and the text itself has been included to cater for other variables which may play a part in the study of language change and variation.

Three of these sub-corpora we intend to present at this conference are *CEPhiT (Corpus of English Philosophy Texts)* and *CELiST (Corpus of English Life Sciences Texts)* and *CHET (Corpus of Historical English Texts).* The first sub-corpus, *CETA* (*Corpus of English Texts on Astronomy*) has already been presented in *ICAME* Conferences in the recent past and will hopefully be released in Spring 2011.

Swaying the medical audience: persuasion in early English medical instructional passages

Martti Sakari Mäkinen
Hanken School of Economics

This paper investigates the argumentation in early science, more specifically the ways in which the readers of early English medical texts were persuaded about the healing properties of charms, remedies, medicines, medical recipes, in other words, the passages written for therapeutic instruction.

The paper sets out to chart the ways in which persuasion is attested in medical instructional passages in the sixteenth and seventeenth century England. The aim is to provide a description of the textual means and strategies used, classified according to the distribution over medical genres and the time of writing. The paper will present a longitudinal study of the linguistic realisations of persuasion, thus also the evolution of persuasive means will be addressed. Eventually, the aim is to find reasons for the observed trends in medical language, possibly in the evolution of medical thinking and the advances in science. It is assumed here that the severe contest of new medical disciplines in the sixteenth and seventeenth centuries brought about such profound changes in scientific and medical ideologies that would have been reflected in the language of science as well, as has been shown to happen.

The paper draws on the speech act theory, in which directives are defined as attempts by the speaker to get the addressee to do something. Other sources of inspiration are historical pragmatics and studies on the evolution of the language of science, corpus linguistics, and genre studies. The study will also consider indirect instruction, thus the question in focus is the difference between the pragmatic meaning and the semantic meaning of a proposition.

The material for the paper will be the corpus of *Early Modern English Medical Texts* (EMEMT; forthcoming 2010), possibly supplemented by material from *Early English Books Online* (EEBO). The corpus contains c. 1.8 million words in 450 medical texts from 1500-1700. The genres or categories, as they are called in the corpus, are general treatises or textbooks, treatises on specific topics, recipe collections and *materia medica*, regimens and health guides, surgical and anatomical treatises, and scientific journals. All of these genres may contain recipes or other instructional passages for the preparation and/or administering of a medicine or a remedy. The estimated number of passages of therapeutic instruction potentially including persuasion is 7,000-10,000.

References

EEBO = *Early English Books Online*. <http://eebo.chadwyck.com/home> Accessed 03/2010.

EMEMT = Corpus of *Early Modern English Medical Texts* forthcoming 2010. Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Ville Marttila, Maura Ratia, Carla Suhr, and Jukka Tyrkkö (eds), with the assistance of Anu Lehto and Alpo Honkapohja. Amsterdam: John Benjamins.

Hamblin, Charles Leonard 1987. *Imperatives*. Oxford: Blackwell.

Hunt, Tony 1990. *Popular Medicine in Thirteenth-Century England: Introduction and Texts*. Cambridge: Brewer.

Jones, Claire 1998. 'Formula and formulation: "Efficacy Phrases" in medieval English medical manuscripts.' *Neuphilologische Mitteilungen* 99:2. 199-210.

Mäkinen, Martti forthcoming 2010. Efficacy phrases in Early Modern English medical recipes'. In Medical Writing in Early Modern English, eds Taavitsainen, Irma and Päivi Pahta. Cambridge: CUP.

Searle, J.R. 1976. 'The classification of illocutionary acts.' *Language in Society*, 5.1: 1-24.

Stannard, Jerry 1982. 'Rezeptliteratur as Fachliteratur'. In Eamon, William (ed.) *Studies on Medieval Fachliteratur*. Brussels: Omirel. 59-73.

Taavitsainen, Irma 2009. 'The pragmatics of knowledge and meaning: Corpus linguistic approaches to changing thought-styles in early modern medical discourse.' In Andreas H. Jucker, Daniel Schreier and Marianne Hundt (eds.) *Corpora: Pragmatics and Discourse, Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)*. (Language and Computers 68.) Amsterdam and New York: Rodopi. 37-62.

Taavitsainen, Irma 2004. 'Genres of Secular Instruction: a linguistic history of useful entertainment.' *Miscelánea: A Journal of English and American Studies* 29. 75-94.

Taavitsainen, Irma, 2001. 'Middle English recipes: Genre characteristics, text type features and underlying traditions of writing.' *Journal of Historical Pragmatics* 2.1: 85-113.

Wear, Andrew, 2000. *Knowledge and Practice in English Medicine, 1550-1680.* Cambridge: Cambridge University Press.

Life is like tennis: serve well and you seldom lose. Issues in building a corpus of American church marquees

Susan Nacey
Hedmark University College

This paper discusses the extent to which it is possible to reconcile theory and practice in the compilation of a corpus of language that is not typically subjected to academic scrutiny: the particularly American form of advertising found in church marquees. Such marquees are outdoor changeable copy or LED signs, typically located on church property but visible to passing motorists. They serve various purposes, which include informing the congregation of church events, boosting the church's attendance and – by extension – finances, and allowing the minister a means of influencing and/or serving the local community. In essence, these signs are billboards for the Lord, one-sentence sermons. Active users change their captions weekly (Rentaria 2005, Shafrir 2007).

Messages on these signs are of academic interest for the study of metaphor, as deliberate use of conventional metaphor (see Steen 2008) provides one effective means of fulfilling church requirements, given the necessarily severe space restrictions of the medium. Documented examples of church messages from how-to books as well as both online and coffee-table photo collections of marquees are replete with pithy captions such as the light-hearted *For a healthy heart, give your faith a workout* or play on words *To prevent sinburn use sonscreen,* the more ominous *Turn or burn*, and the thought-provoking *Aim at nothing and you will always hit it* (Claassen 2005, Glusenkamp 1996, Harvey 2007, Paulson and Paulson 2006, Verbrugge 1999)

To investigate the extent to which such documented collections of church signs reflect actual church practice, a small corpus of weekly photos of church marquees was collected over a four-month period in 2010. Baptist, Methodist, Congregationalist and Pentecostal denominations are represented in the data, together with non-denominational Christian churches.

The present paper focuses on the collection methods employed and the corpus itself, given the lack of consensus over precisely what a collection of language must contain to be deemed a corpus (see e.g. Bowker 2007: 303, Francis 2007: 285, Leech 2007: 316, Tognini-Bonelli 2001: 53). Theoretical considerations such as *representativeness, design criteria, purpose of corpus, length,* and *authenticity* are juxtaposed against the many practical constraints encountered during the collection process of such under-examined, eclectic material.

References

Bowker, Lynne (2007). “Towards a corpus-based approach to terminography.” In *Corpus linguistics: Critical concepts in linguistics*. Vol. 3. W. Teubert and R. Krishnamurthy (eds.). London: Routledge, 303-324.

Claassen, David J. (2005). *Silent words spoken loudly: Church sign sayings*. Lima, OH: CSS Publishing Company, Inc.

Francis, W. Nelson (2007). “Problems of assembling and computerizing large corpora.” In *Corpus linguistics: Critical concepts in linguistics*. Vol. 1. W. Teubert and R. Krishnamurthy (eds.). London: Routledge, 285-298.

Glusenkamp, Ronald T. (1996). *Signs for these times: Church signs that work*. Saint Louis, MO: Concordia Publishing House.

Harvey, L. James (2007). *701 sentence sermons, vol. 4: Attention-getting quotes for church signs, bulletins, newsletters, and sermons*. Grand Rapids, MI: Kregel Publications.

Leech, Geoffrey (2007). “The value of a corpus in English language research.” In *Corpus linguistics: Critical concepts in linguistics*. Vol. 1. W. Teubert and R. Krishnamurthy (eds.). London: Routledge, 315-325.

Paulson, Steve and Pam Paulson (2006). *Church signs across America*. Woodstock, N.Y.: Overlook Press.

Rentaria, Melissa (2005). “Marquees a pulpit for one-sentence sermons.” In *The San Diego Union-Tribune*, Retrieved April 9, 2010 from http://legacy.signonsandiego.com/uniontrib/20050714/news_1c14signs.html.

Shafrir, Doree (2007). “Signs from god: The curious history of church marquees.” In *Slate*, Retrieved April 8, 2010 from http://www.slate.com/id/2167297/.

Steen, Gerard (2008). “The paradox of metaphor: Why we need a three-dimensional model of metaphor.” In *Metaphor & Symbol*, vol. 23 (4), 213-241.

Tognini-Bonelli, Elena (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.

Verbrugge, Verlyn D. (1999). *Time-saving ideas for your church sign: 1001 attention-getting sayings*. Grand Rapids, MI: Zondervan.

Translating English ergative intransitives into Norwegian and Swedish

Lene Nordrum
Chalmers University of Technology

Structures that are in-between active and passive voice have received much attention across linguistic theories. In English, one such construction is exemplified by the second sentence in the alternation pattern: *the sun melted the snow – the snow melted*. The definition of the structure varies according to theory, but will be referred to as ergative intransitive in this paper, following the general theoretical framework of Systemic Functional Linguistics.

Typical for English ergative intransitives is the lack of morphological marking on the verb. In other languages, however, the active-passive role of the ergative intransitive is usually signaled by a reflexive construction or some other overt morphological marking. This lack of formal correspondence between languages can represent a problem in translation.

This paper investigates a number of English ergative verbs and their translations into Norwegian and Swedish. The empirical data are taken from the sister corpora English-Norwegian Parallel Corpus and English-Swedish Parallel Corpus. The direction between English to Norwegian/Swedish is particularly interesting because the two Scandinavian languages have formal correspondences to the English verbs in some cases, but prefer structures with morphological marking in most cases. The goal of the paper is to use the translations to explore the semantic, syntactic and morphological properties of both the English verbs and the verbs used in the translations.

Standardizing the Corpus of Early English Correspondence

Minna Palander-Collin & Mikko Hakala
University of Helsinki

Spelling variation in Early Modern English data poses considerable problems for the accuracy of corpus linguistic tools and methods developed primarily for Modern English. Private writings in particular like personal letters show extensive spelling variation partly reflecting e.g. regional and social differences, but it is not uncommon that a single writer exhibits internal variation and idiosyncratic spellings. Consequently, spelling variation needs to be dealt with in order to improve corpus results in Early Modern English data. To compile a fully standardized corpus is a time-consuming process requiring manual identification of spelling variants, but automatic standardization with a trained program may result in an adequately standardized corpus.

In this poster we shall report on problems and solutions in standardizing the spelling of the seventeenth-century letters in the *Corpus of Early English Correspondence* with the VARD (VARiant Detector) program developed for the purpose of automatic standardization of Early Modern English (see http://www.comp.lancs.ac.uk/~barona/vard2/). Our aim is to test the impact of standardization on results. As the first method of testing, we shall use VARD statistics on variant tokens and types to see how the program works with the data and compare the results with the ones obtained for the *Corpus of Early Modern English Medical Texts* (Lehto, Baron, Ratia & Rayson Forthcoming). We expect this comparison to show whether private writings pose different problems as compared to printed works that can be assumed to be somewhat more standardized than private writings, and whether there is a trend similar to medical texts in our data towards a more uniform spelling during the seventeenth century. Secondly, we shall adopt the end user's perspective and test the WordSmith keyword and cluster tools on the data in order to see how results change depending on the spelling variation of the data. This analysis will include a comparison of manually vs. automatically standardized data.

References

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

Lehto, Anu, Alistair Baron, Maura Ratia & Paul Rayson. Forthcoming/2010. Improving the precision of corpus methods. The standardized version of *Early Modern English Medical Texts*. In: Irma Taavitsainen & Päivi Pahta (eds.), *Early Modern English Medical Texts. Corpus Description and Studies*. Amsterdam/Philadelphia: John Benjamins. 281–344.

New evidence in cognitive processing of formulaic sequences by native speakers and learners of English: An eye-tracking study

Sven Saage, Viktoria Künstler
Justus-Liebig-Universität Giessen

In recent years the interest in the analysis of formulaic language processing by means of eye-tracking has grown largely. Various studies (e.g. Underwood et al. 2004, Siyanova-Chanturia et al. to appear) focused on the question whether and to what degree native and non-native speakers of English process collocations, multiword units, idioms, etc. holistically, with varying and largely inconsistent results. The different results can be pre-eminently traced back to a range of methodological approaches.

The present study is a refined version of an earlier eye-tracking experiment which was conducted last year. In this earlier study, we found some evidence for shorter processing times of formulaic sequences for native speakers compared to proficient and less proficient learners of English. Furthermore, the results indicated that native speakers and non-native speakers alike fixated formulaic sequences fewer and shorter than non-formulaic sequences. The results, however, did not allow for an unambiguous interpretation, as the mean frequencies of the 4-grams and control units deviated too much in some cases.

Building on experience of this earlier study, we have refined our methodology and have created new test sentences which will draw a more distinct picture of the reading of the test sentences and their corresponding control sentences. Analogously to Ellis & Simpson-Vlach's (2009) approach, we extrapolated 4-grams from the BNC, with frequency and mutual information as identification criteria. In the next step, 30 coherent formulaic units were chosen which were then incorporated into small test sentences. For each test sentence we formulated a control sentence, ensuring that each element of the control unit matched the frequency of the items in the formulaic sequence, but did not constitute a frequent 4-gram. Finally, native speakers and learners of English had to read these 60 test sentences on a computer screen while their eyes were being tracked. Currently, we are conducting the statistical analysis of the study.

In our poster we will focus on a detailed description of our methodology. We will also present concrete results from the analysis of our experiment.

References

Ellis, N. C. & Simpson-Vlach, R. (2009): Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory* 5, 61-78.

Siyanova-Chanturia, A., Conklin, K., and Schmitt, N. (in press): Adding more fuel to the fire: An eye-tracking study of idiom processing by native and nonnative speakers. *Second Language Research*.

Underwood G., Schmitt N. & Galpin, A. (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*. Philadelphia: John Benjamins, 153-172.

Wray, A. (2008): *Formulaic Language: Pushing Boundaries.* Oxford: Oxford University Press.

Let's take a look at Celine Dion in action – An empirical investigation of the pragmatic value of Light Verb Constructions

Arian Shahrokny-Prehn
Leibniz University Hannover

The aim of this paper is to provide a usage-based account of stylistic and pragmatic facets of so-called Light Verb Constructions (LVCs), e.g. *take a look*, *have a walk* etc. A number of papers have been published over the decades with varying topics, stemming from very different traditions. One thing that the majority of papers published to date agree on is the 'colloquial flavour' conveyed by LVCs (e.g. Poutsma 1926, Curme 1931, Quirk *et al.* 1985, Wierzbicka 1982). However, empirical data yields a much more complex and heterogeneous picture.

Based on the representation of LVCs in the Corpus of Contemporary American English (COCA), my research shows that they are not at all restricted to informal speech or "highly colloquial" contexts but are indeed moving through the registers. Support for this finding comes from Stein (1991: 26), who suggests that *have a* V expresses "personal courtesy, personal attention and care for another person" plus some higher degree of politeness.

While stylistic constraints are indeed at work (i.e. specific verb classes are preferred by particular genres), a close analysis of selected LVCs will demonstrate that both their distribution and usage are much more versatile than has been acknowledged so far. Especially the perceivable difference in the use of *take a look* with regards to genre will be closely investigated. The 'nature' of the Spoken genre in the COCA, namely that it consists to a large degree of TV transcripts, may play a decisive role here. Although this peculiarity is of benefit for my own research, it has to be taken with quite a grain of salt, since it might also entail that a 400 million word corpus is not as representative as we might want to think.

The research questions guiding through this paper will then be the following:

- What empirical evidence can we find for/against the claim that LVCs are highly colloquial?
- Employing qualitative analysis, can we assert patterns of usage that allow us to draw conclusions regarding the pragmatic/discourse function of LVCs?

Overall, it will be argued that these stylistic constraints and discourse/ pragmatic functions are really at the heart of the issue as to why speakers actually make use of this periphrastic construction when they could just as well use its 'simple verb' counterpart.

References

Curme, George O. 1983 [1931]. *A grammar of the English Language.* Vol.2: *Syntax.* Reprint. Essex, Conn.: Verbatim.

Poutsma, H. 1926. *A Grammar of Late Modern English.* Part II: *Parts of Speech.* Section II: *The Verb and the Particles.* Groningen: P. Noordhoff.

Quirk Randolph *et al.* 1985. *A Comprehensive Grammar of the English Language.* Longman: London.

Stein, Gabriele. 1991. "The phrasal verb type 'to have a look' in modern English". *IRAL* 29, 1-29.

Wierzbicka, Anna. 1982. "Why can you have a drink when you can't *have an eat?" *Language* 58, 753-799.

Representation of context in corpus design: Early Modern English Medical Texts corpus

Jukka Tyrkkö, Raisa Oinonen, Ville Marttila, Turo Hiltunen, Anu Lehto
University of Helsinki

The *Early Modern English Medical Texts (EMEMT)* corpus was published in 2010. With 2 million words and over 450 text samples, EMEMT provides a representative resource for studying the language of Early Modern medicine. The corpus covers a wide variety of texts from the most learned treatises to health guides aimed at literate laymen, and includes a large selection of medical articles from the *Philosophical Transactions of the Royal Society*, the first scholarly periodical in English. Distributed with EMEMT Presenter, a proprietary edition of Corpus Presenter by Raymond Hickey, EMEMT introduces several new and improved features to historical corpora such as a comprehensive catalogue of background information on each text, selected facsimile images from the majority of the texts, and hyperlinks to various online resources — all integrated within the corpus tool and conveniently accessible at any time. In adding these features, the compilers of EMEMT were motivated by a desire to ensure that the philological context be not forgotten, and a belief that contextual information adds valuable depth to corpus linguistic analysis.

The poster will focus on the crucial issue of how contextual information can and should be included in historical corpora. Although corpus linguistic methods rely in the first instance on quantitative data, the realities of historical linguistics dictate that the results cannot be reasonably interpreted without an understanding of the cultural context from which the texts originate. The role of translation, complicated transmission histories, misattribution of texts, and contrasts between the true and assumed background of individual authors will be just some of the issues discussed. The poster will highlight some of these key challenges facing historical corpus linguists, and discuss ways in which contextual data can help explain results and guard against mistakes.

The poster presentation will include a live demonstration of EMEMT.

Alternation of past time reference in World Englishes

Valentin Werner
University of Bamberg

Past time reference represents one of the notorious areas of English grammar. The present perfect in particular seems to defy grammatical categorization and was given various labels such as ‘tense’ (Declerck 2006), ‘aspect’ (Quirk *et al.* 1985), ‘phase’ (Meyer 1992) or ‘status’ (Bauer 1970, cf. Kortmann 1995). Considerable efforts have been invested in establishing theoretical explanations for the usage and alternation of the present perfect in particular and a plethora of factors exerting an influence on the choice between different realizations in typical contexts has been identified. Yet, a consensus among scholars still seems to be out of reach, especially with regard to a semantic description of the present perfect. It has to be noted, however, that in-depth quantitative analyses (Elsness 1997, 2009; Schlüter 2002) could partly contribute to solve the issue.

This is where this on-going project ties in and extends the previous findings beyond the established British versus American English paradigm to a World Englishes perspective. It is based on data from a number of components of the synchronous International Corpus of English (ICE), both from L1-varieties (such as Irish or New Zealand English) and L2-varieties (such as Singapore or Philippine English). The aim is to determine the weight of different factors within and outside the verb phrase as to the choice between present perfect forms and other variants of past time reference. A statistical analysis would enable us to establish a measure of similarity between the different components of the corpus and to determine if the grammatical area of past time reference lies within the common core of Englishes or if varieties, genres and text types pattern rather idiosyncratically. In the end, possibly under consideration of extralinguistic factors such as colonial history and language policy, we could come to more general conclusions with regard to the adequacy of global theories on World Englishes (such as the Kachruvian Circle Model, Modiano 1999 or Schneider 2007) for the grammatical area under consideration.

ARCHER past and present (1990-2010)

Nuria Yanez-Bouza
University of Manchester

The purpose of this paper is to provide an up-to-date account of ARCHER, *A Representative Corpus of Historical English Registers*, twenty years after it was first compiled by Douglas Biber and Edward Finegan in the early 1990s.

ARCHER is a multi-genre historical corpus of British and American English covering the period 1650-1999, ca. 1.8 million words. It exists in three versions and is on its way towards the fourth one: ARCHER-1 (1990-93), ARCHER-2 (2004-05), ARCHER 3.1 (2006), and ARCHER 3.2 (forthcoming). Although the corpus has been used in a large number of studies in the history of English, and the number and geographical locations of the consortium have enlarged over the years, a comprehensive description of its design and contents since its early days is still lacking (cf. Biber *et al.* 1994a, 1994b; Biber and Finegan 1997: 255-257). In order to fill the gap, this paper will tell the history of ARCHER from 1990 to 2010.

The approach is two-fold: past and present. I will first describe the original aims and design of ARCHER-1 along with the additions in ARCHER-2. In doing so, I will bring to light the (hitherto unnoticed) existence of slightly divergent versions of the corpus. The focus will then lie in the current version, ARCHER 3.1, including (i) an account of the changes in its structure and contents compared to earlier versions; (ii) a summary of the new coding conventions for text annotation and of the edits carried out, primarily with regard to filenames; and (iii) remarks on access to the corpus. The paper will end with an outlook of the ongoing phase towards ARCHER 3.2: aims, work in progress and expectations.

References

ARCHER website. http://llc.stage.manchester.ac.uk/research/projects/archer/.

Biber, Douglas and Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In Terttu Nevalainen and Leena Kahlas-Tarkka (eds.), *To explain the present*, 253-275. Helsinki: Société Néophilologique.

Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994a. *ARCHER* and its challenges: Compiling and exploring *A Representative Corpus of Historical English Registers*. In Udo Fries, Peter Schneider and GottieTottie (eds.), *Creating and using English language corpora*, 1-13. Amsterdam: Rodopi.

Biber, Douglas, Edward Finegan, Dwight Atkinson, Ann Beck, Dennis Burges and Jena Burges. 1994b. The design and analysis of the *ARCHER* corpus: A progress report [*A Representative Corpus of Historical English Registers*]. In Merja Kytö, Matti Rissanen and Susan Wright (eds.), *Corpora across the centuries*, 3-6. Amsterdam and Atlanta: Rodopi.

List of participants

Aijmer	Karin	University of Gothenburg	karin.aijmer@eng.gu.se
Alsop	Sian	Coventry University	alsops@uni.coventry.ac.uk
Altenberg	Bengt	Lund University	fam.altenberg@telia.com
Andersen	Gisle	Norwegian School of Economics and Business Administration (NHH)	gisle.andersen@nhh.no
Anglemark	Linnéa	University of Uppsala	Linnea.Anglemark@engelska.uu.se
Arús-Hita	Jorge	Universidad Complutense de Madrid	jarus@filol.ucm.es
Axelsson	Karin	University of Gothenburg	karin.axelsson@eng.gu.se
Aydin	Ada Benedicte	University of Oslo	adaba@student.ilos.uio.no
Azevedo	Flávia	Federal University of Santa Catarina	flviaazevedo@yahoo.com.br
Bachmann	Ingo	University of Duisburg-Essen	ingo.bachmann@uni-due.de
Baron	Alistair	Lancaster University	a.baron@comp.lancs.ac.uk
Bartsch	Sabine	Technische Universität Darmstadt	bartsch@linglit.tu-darmstadt.de
Bech	Kristin	University of Oslo	kristin.bech@ilos.uio.no
Berglund-Prytz	Ylva	University of Oxford	ylva.berglund@oucs.ox.ac.uk
Bernaisch	Tobias	Justus Liebig University Giessen	Tobias.J.Bernaisch@anglistik.uni-giessen.de
Bernardini	Silvia	University of Bologna	silvia.bernardini@unibo.it
Biber	Doug	Northern Arizona University	douglas.biber@nau.edu
Biewer	Carolin	University of Zurich	carolin.biewer@es.uzh.ch
Blackwell	Susan	University of Birmingham	s.a.blackwell@bham.ac.uk
Blanco-Suárez	Zeltia	University of Santiago de Compostela	zeltia.blanco@rai.usc.es
Borchers	Melanie	University of Duisburg-Essen	melanie.borchers@uni-due.de
Breban	Tine	University of Leuven	tine.breban@arts.kuleuven.be
Bredenbröcker	Martina	University of Potsdam	martina.bredenbroecker@uni-potsdam.de
Brems	Lot	University of Leuven	lieselotte.brems@arts.kuleuven.be
Brown	Mark	Norwegian School of Business BI	mark.brown@bi.no
Börjesson	Viktoria	University of Gothenburg	viktoria.borjesson@eng.gu.se

Callies	Marcus	Johannes-Gutenberg-Universität Mainz	mcallies@uni-mainz.de
Cao	Jing-Xiang	Dalian university of Technology; Lancaster University	alicia1973@yahoo.cn
Claridge	Claudia	University of Duisburg-Essen	claudia.claridge@uni-due.de
Coffey	Stephen James	Università di Pisa	coffey@cli.unipi.it
Collins	Peter	University of New South Wales	p.collins@unsw.edu.au
Crespo	Begoña	University of A Coruña	bcrespo@udc.es
Crystal	David	University of Bangor	davidcrystal1@googlemail.com
Curzan	Anne	University of Michigan	acurzan@umich.edu
Dant	Doris	Brigham Young University	drdant@gmail.com
Davidse	Kristin	K.U.Leuven (University of Leuven)	kristin.davidse@arts.kuleuven.be
De Cock	Sylvie	Université catholique de Louvain	sylvie.decock@uclouvain.be
De Felice	Rachele	University of Nottingham	rachele.de_felice@nottingham.ac.uk
Diemer	Stefan	Technical University Berlin	s.diemer@umwelt-campus.de
Dose	Stefanie	Justus Liebig University Giessen	Stefanie.Dose@anglistik.uni-giessen.de
Dossena	Marina	Università degli Studi di Bergamo	marina.dossena@unibg.it
Ebeling	Jarle	University of Oslo	jarle.ebeling@usit.uio.no
Ebeling	Signe Oksefjell	University of Oslo	s.o.ebeling@ilos.uio.no
Egan	Thomas	Hedmark University College	Thomas.Egan@hihm.no
Elgemark	Anna	Göteborgs Universitet	anna.elgemark@eng.gu.se
Elsness	Johan	University of Oslo	johan.elsness@ilos.uio.no
Fabricius-Hansen	Cathrine	University of Oslo	c.f.hansen@ilos.uio.no
Facchinetti	Carla	Secondary School – Lyceum	carla.facchinetti@istruzione.it
Facchinetti	Roberta	University of Verona	roberta.facchinetti@univr.it
Fanego	Teresa	University of Santiago de Compostela	teresa.fanego@usc.es
Ferraresi	Adriano	University of Naples “Federico II” / University of Bologna	adriano@sslmit.unibo.it
Fryer	Daniel Lees	University of Gothenburg, Sweden	daniel.lees.fryer@sprak.gu.se
Fuoli	Matteo	Università di Trento	matteo.fuoli@gmail.com
Garretson	Gregory	Uppsala University	gregory.garretson@engelska.uu.se
Gee	Matthew	Birmingham City University	matt.gee@bcu.ac.uk

Gonzalez-Diaz	Victorina	University of Liverpool	vgdiaz@liv.ac.uk
Goossens	Diane	Université catholique de Louvain	diane.goossens@uclouvain.be
Graedler	Anne-Line	Hedmark University College	anneline.graedler@hihm.no
Granger	Sylviane	Université catholique de Louvain	sylviane.granger@uclouvain.be
Grant	Lynn	Auckland University of Technology	lynn.grant@aut.ac.nz
Groom	Nicholas	University of Birmingham	n.w.groom@bham.ac.uk
Götz	Sandra	Justus Liebig University, Giessen & Macquarie University, Sydney	Sandra.Goetz@anglistik.uni-giessen.de
Hakala	Mikko	University of Helsinki	mikko.hakala@helsinki.fi
Halbe	Dorothea	Univerity of Trier	halbed@uni-trier.de
Hardie	Andrew	Lancaster University	a.hardie@lancaster.ac.uk
Hasselgård	Hilde	University of Oslo	hilde.hasselgard@ilos.uio.no
He	An E	The Hong Kong Institute of Education	heane@ied.edu.hk
Heslien	Siri	University of Oslo	saheslie@student.uio.no
Hoehn	Nicole	University of Basel	nicole.hoehn@unibas.ch
Hoffmann	Sebastian	University of Trier	hoffmann@uni-trier.de
Hofland	Knut	Uni Research AS	knut.hofland@uni.no
Huber	Magnus	University of Giessen	magnus.huber@anglistik.uni-giessen.de
Izquierdo	Marlén	University of León	marlen.izquierdo@unileon.es
Johannessen	Janne Bondi	University of Oslo	j.b.johannessen@iln.uio.no
Johansen	Torunn	University of Oslo	torunnnj@student.ilos.uio.no
Kaltenböck	Gunther	University of Vienna	gunther.kaltenboeck@univie.ac.at
Kehoe	Andrew	Birmingham City University	andrew.kehoe@bcu.ac.uk
Kirk	John	Queen’s University Belfast	j.m.kirk@qub.ac.uk
Kobayashi	Yuichiro	University of Osaka	kobayashi0721@gmail.com
Koch	Christopher	Justus Liebig University, Giessen	christopher.koch@anglistik.uni-giessen.de
Kohnen	Thomas	University of Cologne	thomas.kohnen@uni-koeln.de
Kunz	Kerstin Anna	Saarland University	k.kunz@mx.uni-saarland.de
Künstler	Viktoria	Justus-Liebig-Universität Giessen	viktoria.kuenstler@zmi.uni-giessen.de
Kytö	Merja	Uppsala University	merja.kyto@engelska.uu.se

Labrador	Belén	University of León	belen.labrador@unileon.es
Laing	Margaret	University of Edinburgh	m.laing@ed.ac.uk
Lareo	Inés	University of A Coruña	ilareo@udc.es
Larsson Aas	Hege	University of Oslo	Hege.larssonaas@gmail.com
Lavid	Julia	Universidad Complute	lavid@filol.ucm.es
Leedham	Maria	The Open University	m.e.leedham@open.ac.uk
Lefer	Marie-Aude	Université catholique de Louvain	marie-aude.lefer@uclouvain.be
Lehmann	Hans Martin	University of Zurich	hmlehman@es.uzh.ch
Levin	Magnus	Linnaeus University	magnus.levin@lnu.se
Lindmark	Kerstin	Stockholm University	kerstin.lindmark@ling.su.se
Lindquist	Hans	Malmö University	hans.lindquist@mah.se
Littré	Damien	Université Catholique de Louvain	damien.littre@uclouvain.be
Liu	Anne Li-E	University of Nottingham	aexlel1@nottingham.ac.uk
Ljung	Magnus	Stockholm University	Magnus.Ljung@English.su.se
López-Couso	María José	University of Santiago de Compostela	mjlopez.couso@usc.es
Lorenz	David	Universität Freiburg	david.lorenz@frequenz.uni-freiburg.de
Louw	William	University of Zimbabwe	louwbill@gmail.com
Lozano	Cristobal	Universidad de Granada	cristoballozano@ugr.es
Luckmann	Kathrin	University of Duisburg-Essen	kathrin.luckmann@uni-due.de
Macháček	Jaroslav	Palacký University	jaroslav.machacek@upol.cz
Mair	Christian	University of Freiburg	christian.mair@anglistik.uni-freiburg.de
Majewski	Stefan	University of Vienna	stefan.majewski@univie.ac.at
Martinkova	Michaela	Palacký University	michaela.martinkova@upol.cz
Meijs	Willem	Language Consultancy Desk Birmingham	wjmeijs@gmail.com
Mendikoetxea	Amaya	UNIVERSIDAD AUTÓNOMA DE MADRID	amaya.mendikoetxea@uam.es
Meunier	Fanny	Université Catholique de Louvain	fanny.meunier@uclouvain.be
Meyer	Charles	University of Massachusetts Boston	Meyer@cs.umb.edu
Mindt	Ilka	University of Potsdam	ilka.mindt@uni-potsdam.de
Moessner	Lilo	RWTH University Aachen	moessner@anglistik.rwth-aachen.de

Mohamed	Ghada	Lancaster University	g.mohammed@lancaster.ac.uk
Mollin	Sandra	University of Heidelberg	sandra.mollin@as.uni-heidelberg.de
Moskowich	Isabel	University of A Coruña	imoskowich@udc.es
Mukherjee	Joybrato	Justus Liebig University Giessen	mukherjee@uni-giessen.de
Mäkinen	Martti	Hanken School of Economics	martti.makinen@hanken.fi
Nacey	Susan	Hedmark University College	susan.nacey@hihm.no
Neuhaus	Elisabeth Maria	University of Oslo	elisabeth.neuhaus@gmail.com
Nevalainen	Terttu	University of Helsinki	terttu.nevalainen@helsinki.fi
Nordrum	Lene	Chalmers University of Technology	lene.nordrum@chalmers.se
Nurmi	Arja	University of Helsinki	arja.nurmi@helsinki.fi
Oinonen	Raisa	University of Helsinki	raisa.oinonen@helsinki.fi
Opdahl	Lise	University of Bergen	lise.opdahl@if.uib.no
Osimk	Ruth	University of Vienna	ruth.osimk@univie.ac.at
Palander-Collin	Minna	University of Helsinki	minna.palander-collin@helsinki.fi
Pérez-Blanco	María	Universidad de León	mperb@unileon.es
Peters	Pam	Macquarie University	pam.peters@mq.edu.au
Peukert	Hagen	University of Hamburg	hagen.peukert@uni-kassel.de
Rabadán	Rosa	University of León	rosa.rabadan@unileon.es
Radeka	Michael	University of Vienna	michael.radeka@univie.ac.at
Rayson	Paul	Lancaster University	p.rayson@lancs.ac.uk
Reichardt	Renate	University of Birmingham	renate_reichardt@hotmail.com
Renouf	Antoinette	Birmingham City University	ajrenouf@bcu.ac.uk
Reppen	Randi	Northern Arizona University	randi.reppen@nau.edu
Rissanen	Matti	University of Helsinki	matti.rissanen@helsinki.fi
Roethlisberger	Melanie	University of Zurich	melanie.roethlisberger@uzh.ch
Rogatcheva	Svetla	Justus-Liebig Universität Giessen	svetlomira.i.rogatcheva@anglistik.uni-giessen.de
Rohdenburg	Günter	University of Paderborn	rohdenburg@onlinehome.de
Ruetten	Tanja	University of Cologne	tanja.ruetten@uni-koeln.de
Rørvik	Sylvi	Hedmark University College	sylvi.rorvik@hihm.no

Saage	Sven	Justus-Liebig Universität Giessen	sven.saage@zmi.uni-giessen.de
Santos	Diana	University of Oslo	d.s.m.santos@ilos.uio.no
Schmidt	Christa Maria	RWTH University Aachen	cschmidt@fb7.rwth-aachen.de
Schneider	Gerold	University of Zurich	gschneid@es.uzh.ch
Schulz	Monika Edith	Hamburg University	monika-edith.schulz@uni-hamburg.de
Schutz	Natassia	Université Catholique de Louvain	natassia.schutz@uclouvain.be
Shahrokny-Prehn	Arian	Leibniz Universität Hannover	arian.shahrokny@engsem.uni-hannover.de
Smith	Adam	Macquarie University	adam.smith@mq.edu.au
Smitterberg	Erik	Uppsala University	erik.smitterberg@engelska.uu.se
Sotillo	Susana	Montclair State University	sotillos@mail.montclair.edu
Spina	Stefania	Università per Stranieri Perugia	stefania.spina@unistrapg.it
Stenbrenden	Gjertrud F.	University of Oslo	g.f.stenbrenden@ilos.uio.no
Stenström	Anna-Brita	Bergen University	ab.stenstrom@telia.com
Stubbs	Michael	University of Trier	stubbs@uni-trier.de
Suarez-Gomez	Cristina	University of the Balearic Islands	cristina.suarez@uib.es
Sudicky	Petr	Masaryk University	sudicky@phil.muni.cz
Svartvik	Jan	Lund University	jan.svartvik@telia.com
Taavitsainen	Irma	University of Helsinki	irma.taavitsainen@helsinki.fi
Tagg	Caroline	Open University	c.tagg@open.ac.uk
Tagliamonte	Sali	University of Toronto	sali.tagliamonte@utoronto.ca
Teich	Elke	Universitaet des Saarlandes	e.teich@mx.uni-saarland.de
Thunes	Martha	University of Bergen	martha.thunes@lle.uib.no
Tognini Bonelli	Elena	University of Siena	elena@twc.it
Tottie	Gunnel	University of Zurich	gtottie@mac.com
Trotta	Joe	University of Gothenberg	joe.trotta@eng.gu.se
Tyrkkö	Jukka	University of Helsinki	jukka.tyrkko@helsinki.fi
Uhrig	Peter	Universität Erlangen-Nürnberg	peter.uhrig@angl.phil.uni-erlangen.de
Vaes	Kees	John Benjamins Publishing	kees.vaes@benjamins.nl
Vartiainen	Turo	University of Helsinki	turo.vartiainen@helsinki.fi

Velupillai	Viveka	Giessen	viveka.velupillai@email.de
Vetchinnikova	Svetlana	University of Helsinki	svetlana.vetchinnikova@helsinki.fi
Viberg	Åke	Uppsala Universitet	Ake.Viberg@lingfil.uu.se
Vincent	Benet	University of Birmingham	BDV700@bham.ac.uk
Werner	Valentin	University of Bamberg	valentin.werner@uni-bamberg.de
Wikberg	Kay	University of Oslo	kay.wikberg@ilos.uio.no
Wynne	Martin	University of Oxford	martin.wynne@oucs.ox.ac.uk
Yamazaki	Shunji	Daito Bunka University	yamazaki@ic.daito.ac.jp
Yanez-Bouza	Nuria	University of Manchester	nuria.yanez-bouza@manchester.ac.uk
Yang	Suxiang	Henan Polytechnic University	jzysx@126.com
Yao	Xinyue	University of New South Wales	xinyue.yao@unsw.edu.au
Zaytseva	Ekaterina	Johannes Gutenberg University Mainz	zaytseve@uni-mainz.de